Geophysical Journal International

Geophys. J. Int. (2018) **215,** 1003–1021
Advance Access publication 2018 July 25
GJI Seismology

doi: 10.1093/gji/ggy303

Optimal regularization for a class of linear inverse problem

Andrew P. Valentine and Malcolm Sambridge

Research School of Earth Sciences, The Australian National University, Canberra ACT 2601, Australia. E-mail: andrew.valentine@anu.edu.au

Accepted 2018 July 25. Received 2018 July 4; in original form 2018 March 18

SUMMARY

Most linear inverse problems require regularization to ensure that robust and meaningful solutions can be found. Typically, Tikhonov-style regularization is used, whereby a preference is expressed for models that are somehow 'small' and/or 'smooth'. The strength of such preferences is expressed through one or more damping parameters, which control the character of the solution, and which must be set by the user. However, identifying appropriate values is often regarded as a matter of art, guided by various heuristics. As a result, such choices have often been the source of controversy and concern. By treating these as hyperparameters within a hierarchical Bayesian framework, we are able to obtain solutions that encompass the range of permissible regularization parameters. Furthermore, we show that these solutions are often well-approximated by those obtained via standard analysis using certain regularization choices which are—in a certain sense—optimal. We obtain algorithms for determining these optimal values in various cases of common interest, and show that they generate solutions with a number of attractive properties. A reference implementation of these algorithms, written in Python, accompanies this paper.

Key words: Inverse theory; Probability distributions; Statistical methods.

1 INTRODUCTION

Linear—or linearized—inverse problems are commonly encountered in geophysics and beyond, and the starting point for their solution is usually the least-squares algorithm. Since the advent of computational tools for linear algebra, some variant of this approach has been used in a huge variety of settings where it is desired to fit models to data: examples include imaging of the Earth's interior at all scales (e.g. Woodhouse & Dziewonski 1984; Pratt & Worthington 1990; Spakman *et al.* 1993; Lekić & Romanowicz 2011; Ritsema *et al.* 2011; Deuss *et al.* 2013); systematic and targeted earthquake analyses (e.g. Dziewonski *et al.* 1981; Bernardi *et al.* 2004; Yagi & Fukahata 2011); models for the Earth's geoid and derived quantities (e.g. Pavlis *et al.* 2012; Hoggard *et al.* 2016); and much more. However, for many realistic inversion scenarios, there is insufficient information in the data set to constrain all model parameters independently. As a result, it is routine to introduce some form of regularization—or, from a statistician's perspective, prior information—ensuring that a unique solution to the least-squares problem exists.

A perennial question therefore arises: what is the appropriate regularization to use in a given problem? Typically, solutions of a certain 'style' can be preferred on physical or philosophical grounds: thus, regularizers that promote smooth or small models are commonly encountered. However, it remains necessary to determine the weights assigned to regularization terms (often known as 'damping parameters'): to what extent should smoothness (for example) be preferred at the expense of fitting fine-scale features in the data? In general, no principled way to address such questions has been found. Instead, the damping parameters are often chosen in a rather *ad hoc* fashion, guided by various heuristics. As a result, regularization choices are a common source of confusion and controversy: is a model providing meaningful insights, or merely reflecting the authors' choices? For example, Boschi & Dziewoński (1999) identify regularization as one of the principal reasons for discrepancies between the tomographic models produced by different research groups—an issue that persists almost two decades later, as shown in Schaeffer & Lebedev (2015).

Most of the existing strategies for guiding regularization choices are developed from principles of numerical analysis and linear algebra. In this paper, we instead approach the issue from a hierarchical Bayesian perspective. This builds on an idea first introduced—in a different context—by Mackay (1992a,b), and is an extension of the Bayesian approach introduced into geophysics by Tarantola & Valette (1982). This allows us to introduce a natural strategy for defining and identifying 'optimal' regularization parameters, and the results turn out to have meaningful interpretation even in the context of non-Bayesian inverse theory. However, the Bayesian approach brings numerous advantages, including the ability to then 'integrate out' the regularization parameters to obtain a posterior distribution that encapsulates any uncertainties due to regularization choices.

In principle, the resulting algorithm can be entirely free from any user-determined parameters. However, this relies on the inverse problem having 'nice' properties: the various assumptions underpinning linear inversion must be met (including linearity of the forward model, and data noise with known, Gaussian form), and the data must illuminate the model parameters in an unbiased fashion. In practice, these conditions are rarely satisfied. To mitigate this, the user may express preferences regarding the desired level of regularization, in the form of a 'hyperprior' on the regularization parameters. However, such preferences may be relatively weak, and the solution is less dependent on such qualitative choices than in a conventional approach to regularization.

Conceptually, the procedure we develop operates in two stages: first, some statistical properties of the desired solution are estimated from the data; then, the inversion is constrained so as to meet these criteria. The implementation of this approach depends upon the extent to which the user is prepared to impose a pre-determined covariance structure onto the model parameters. In the most general case, a nonlinear search must be performed to identify a stable solution to two coupled systems. This is computationally expensive, but nevertheless may be tractable for problems of modest model dimension, and a reference Python implementation is provided as a supplement to this paper. However, in the common case of 'Tikhonov' regularization, where a covariance structure is imposed *a priori*, the algorithm may be reduced to a root-finding problem in a spectral domain, and computational costs are similar to those of 'traditional' inversion strategies. Again, reference Python implementations are provided.

Our focus in this paper is only on linear inverse problems. Although non-linear inverse problems are frequently tackled through an iterative application of linearized theory, this poses some challenges for our approach as certain assumptions may no longer hold. A full treatment needs to account for the complex propagation of prior information over multiple iterations (see Valentine & Trampert 2016). We discuss such difficulties briefly at the end of this paper, and suggest some potential routes forward. However, a comprehensive treatment must await further work.

2 THEORETICAL DEVELOPMENT

We begin by setting out a brief summary of the standard framework of linear inverse theory, in order to establish definitions and notations that will be required elsewhere in this paper. Inverse theory can be approached from two perspectives: 'deterministic', which is developed based on principles of linear algebra and optimization, and 'probabilistic', where the inverse problem and its solution are framed in terms of probability density functions. For present purposes, it will be useful to review both. We also briefly discuss some common existing strategies for determination of regularization parameters, before setting out our proposed approach.

2.1 The 'deterministic' approach

In a linear inverse problem, we assume that the N-dimensional observed data, $\mathbf{d_0}$, arise as a linear combination of the M model parameters, $\mathbf{m}$. We therefore write

$$\mathbf{d_0} = \mathbf{G}\mathbf{m}\,, \tag{1}$$

with $\mathbf{G}$ representing an $N \times M$ matrix describing the forward operator. In typical geophysical problems, we have $N \gg M$, and poorly conditioned $\mathbf{G}$. Since $\mathbf{d_0}$ is invariably contaminated by observational noise, we do not expect eq. (1) to have an exact solution. Instead, we quantify the extent to which $\mathbf{m}$ can explain the data via the squared L_2 norm of the residuals,

$$\phi(\mathbf{m}) = \|\mathbf{d_0} - \mathbf{G}\mathbf{m}\|_2^2\,, \tag{2}$$

and we minimize this quantity with respect to the elements of $\mathbf{m}$. This procedure, which is described in detail in standard texts such as Menke (1989), leads to the well-known least-squares estimation formula,

$$\mathbf{m} = \left(\mathbf{G^T G}\right)^{-1}\mathbf{G^T d_0}\,. \tag{3}$$

However, when $\mathbf{G}$ is poorly conditioned, this inverse cannot be found stably. From a physical perspective, this implies that the naïve solution of eq. (3) will yield a solution that is unduly sensitive to noise-level variations in the data vector, and which cannot therefore be regarded as robust.

An equivalent statement of this difficulty is that, in many circumstances, the objective function $\phi(\mathbf{m})$ defined by eq. (2) does not have a well-defined unique minimum; instead, there are many models that all explain the data with a similar degree of accuracy. Regularization aims to avoid this situation, via modification of ϕ. Typically, this is achieved by addition of a quadratic term,

$$\phi(\mathbf{m}) = \|\mathbf{d} - \mathbf{G}\mathbf{m}\|_2^2 + \epsilon^2\|\mathbf{D}\mathbf{m}\|_2^2\,, \tag{4}$$

where ϵ^2 is a positive constant, and $\mathbf{D}$ is some pre-determined matrix, chosen to serve as a metric for the 'complexity' of the solution (in a sense appropriate to the problem at hand). It is then straightforward to show that (again, see Menke 1989)

$$\mathbf{m} = \left(\mathbf{G^T G} + \epsilon^2\mathbf{D^T D}\right)^{-1}\mathbf{G^T d_0}\,. \tag{5}$$

With appropriate choices for $\mathbf{D}$ and ϵ^2, a stable inverse can be found. Most commonly, a form known as 'Tikhonov' regularization is adopted. It is convenient to express this as

$$\epsilon^2 \mathbf{D}^{\mathbf{T}}\mathbf{D} = \alpha^2 \mathbf{I} + \beta^2 \mathbf{H}\,, \tag{6}$$

where $\mathbf{I}$ represents the $M \times M$ identity matrix, and $\mathbf{H}$ is some symmetric operator, often chosen to promote 'smooth' solutions to the inverse problem. Often, the second term is omitted (i.e. $\beta = 0$): sometimes this is referred to as 'zeroth order' Tikhonov regularization, or as 'ridge regression', particularly in the literature of numerical analysis.

2.2 Current strategies for choosing regularization parameters

To use regularized least squares, we therefore need to specify α and β (or, more generally, the elements of $\mathbf{D}^{\mathbf{T}}\mathbf{D}$). However, there is no clear strategy for selecting these parameters. Perhaps the most common approach (at least in the geophysical literature) involves exploring the tradeoff between the recovered model norm, $\mathbf{m}^{\mathbf{T}}\mathbf{m}$, and the residuals, $(\mathbf{d_0} - \mathbf{Gm})^{\mathbf{T}}(\mathbf{d_0} - \mathbf{Gm})$, as the regularization is varied. Heavy regularization results in a 'small' (low-norm) model, but a poor ability to explain observations; less regularization improves the data fit, at the expense of a larger model.

Commonly, if only a single regularization parameter needs to be determined (typically the α of ridge regression), the norms of model and residuals are plotted against one another, to give an 'L-curve' (e.g. Hansen 1992). This name comes from the curve's characteristic shape, and the preferred regularization parameter is then chosen by identifying the 'elbow' of the curve. The strategy is justified based on the principle of Occam's razor, which advocates reliance on the simplest (in the present context, smallest) model that can explain observations. Of course, identifying this point is inherently subjective, and can be influenced by immaterial factors such as the scales chosen for plotting (e.g. Constable *et al.* 2015). Thus, regularization choices can be the source of controversy, with argument over whether more- or less-conservative choices could be made that might alter the interpretation of results. This subjectivity also creates difficulties when it is desired to compare results from different inversions (perhaps with distinct data sets or based on alternative methodological choices): how can one ensure that regularization choices are 'equivalent' across multiple examples? Various efforts have been made to introduce more rigour, for example by analysis of the curvature of the L-curve (e.g. Hansen & O'Leary 1993), but these are relatively rarely used. In any case, the existence of a discernible elbow is not guaranteed, and may depend on factors such as the appropriateness of the model parametrization chosen for inversion.

Many other strategies for selecting regularization parameters exist, and some are reviewed in standard texts (e.g. Aster *et al.* 2013). These include Morozov's 'discrepancy principle' (sometimes translated as 'error principle'; Morozov 1968), whereby a recovered model is sought that leaves residuals equivalent to the expected data uncertainty, and techniques based on quantifying model stability and predictive power when only a subset of observations are used, such as PRESS (Allen 1974) or Generalized Cross-Validation (e.g. Golub *et al.* 1979), and even dynamically varying regularization (Rawlinson *et al.* 2008). However, our experience is that these are less widespread within the geophysical community than the L-curve approach.

2.3 The Bayesian perspective

The results of Section 2.1 are derived within a purely algebraic framework. However, they can be endowed with a Bayesian interpretation, as described in the work of Tarantola & Valette (1982). The starting point for Bayesian analysis involves specifying a prior distribution over the model space: that is, we must describe our state of knowledge about $\mathbf{m}$ before the data set $\mathbf{d_0}$ was obtained and analysed. If we assume that this prior is Gaussian in form, centred upon some point $\mathbf{m_p}$ and with covariance matrix $\mathbf{C_m}$, we can write

$$\mathbb{P}[\mathbf{m}] = \frac{1}{\sqrt{(2\pi)^M\,|\mathbf{C_m}|}} \exp\left[-\frac{1}{2}\left(\mathbf{m}-\mathbf{m_p}\right)^{\mathbf{T}}\mathbf{C_m^{-1}}\left(\mathbf{m}-\mathbf{m_p}\right)\right]. \tag{7}$$

We introduce a further assumption: that observational noise can be modelled by a Gaussian with covariance matrix $\mathbf{C_d}$, centred upon the predictions of the linear theory (eq. 1), so that the conditional probability that a model $\mathbf{m}$ generates observations $\mathbf{d}$ is given by

$$\mathbb{P}[\mathbf{d}\,|\,\mathbf{m}] = \frac{1}{\sqrt{(2\pi)^N\,|\mathbf{C_d}|}} \exp\left[-\frac{1}{2}\left(\mathbf{d}-\mathbf{Gm}\right)^{\mathbf{T}}\mathbf{C_d^{-1}}\left(\mathbf{d}-\mathbf{Gm}\right)\right]. \tag{8}$$

Applying Bayes' theorem (Bayes 1763), it is then possible to show that for our specific set of observations, $\mathbf{d_0}$, the posterior distribution is given by

$$\mathbb{P}[\mathbf{m}\,|\,\mathbf{d_0}] = \frac{1}{\sqrt{(2\pi)^N\,|\mathbf{\Psi}|}} \exp\left[-\frac{1}{2}\left(\mathbf{m}-\mathbf{m}_{\text{est}}\right)^{\mathbf{T}}\mathbf{\Psi^{-1}}\left(\mathbf{m}-\mathbf{m}_{\text{est}}\right)\right], \tag{9a}$$

where

$$\mathbf{m}_{\text{est}} = \mathbf{m_p} + \left(\mathbf{G}^{\mathbf{T}}\mathbf{C_d^{-1}}\mathbf{G} + \mathbf{C_m^{-1}}\right)^{-1}\mathbf{G}^{\mathbf{T}}\mathbf{C_d^{-1}}\left(\mathbf{d_0} - \mathbf{Gm_p}\right) \tag{9b}$$

describes the mean of the distribution—and hence represents the most likely solution to the inverse problem—with

$$\mathbf{\Psi} = \left(\mathbf{G}^{\mathbf{T}}\mathbf{C_d^{-1}}\mathbf{G} + \mathbf{C_m^{-1}}\right)^{-1} \tag{9c}$$

describing its covariance. Comparing this with the deterministic solution, eq. (5), the similarities are readily apparent. Although the noise covariance matrix does not feature in eq. (5), it can straightforwardly be introduced as a re-weighting of individual data points. From a Bayesian perspective, choosing regularization parameters is equivalent to identifying the covariance matrix of a Gaussian model-space prior. Thus, any question regarding the 'correct' regularization strategy to use for deterministic inversion has a direct counterpart regarding the 'correct' prior to use in a Bayesian setting. However, as we shall show, the second question is perhaps more readily addressed than the first.

2.4 Building a sensible prior: a hierarchical Bayesian approach to regularization

The model-space prior of eq. (7) is specified by three essential features: it takes the form of a known (Gaussian) class of functions; it is centred on a particular point ($\mathbf{m_p}$) in model-space; and its scale-length in any direction is governed by the covariance matrix $\mathbf{C_m}$. Of these, the first is an assertion, chosen largely for analytical convenience—few other choices of prior admit algebraic solutions to the inverse problem. The second is, generally, an expression of concrete prior knowledge: a reasonable value of $\mathbf{m_p}$ can be identified from pre-existing theory or analysis. However, there is often little to guide the specification of the width (i.e. covariance matrix) of the prior. Indeed, rather than being based on rigorous assessment of the relative probability of different circumstances, this is often treated as something the researcher should tune until results are deemed satisfactory. The hierarchical Bayesian approach aims to introduce more rigour into this process.

Up to this point, our discussion of inverse theory has been entirely standard. We now introduce the modification which lies at the heart of this paper, inspired by the work of Mackay (1992a,b) on backpropagation learning in neural networks. We revise our statement of prior knowledge to be more consistent with what we know, or are willing to assume: the prior continues to have Gaussian form, and be centred upon $\mathbf{m_p}$, but the covariance of this Gaussian is not known. In some cases, we may be willing to assume a certain structure for the covariance matrix: thus, in general, we can introduce a parametric covariance matrix $\mathbf{C_m}(\boldsymbol{\xi})$, where $\boldsymbol{\xi}$ is a vector encapsulating some number, K, of free parameters ($1 \leq K \leq M(M+1)/2$, since a valid covariance matrix must be symmetric positive-definite). The value of $\boldsymbol{\xi}$ is an unknown, as is $\mathbf{m}$: in order to solve the inverse problem, both must be determined. Thus, a complete prior for our inference task is given by

$$\mathbb{P}[\mathbf{m}, \boldsymbol{\xi}] = \mathbb{P}[\mathbf{m} \mid \boldsymbol{\xi}]\,\mathbb{P}[\boldsymbol{\xi}] = \frac{\mathbb{P}[\boldsymbol{\xi}]}{\sqrt{(2\pi)^M\,|\mathbf{C_m}(\boldsymbol{\xi})|}} \exp\left[-\frac{1}{2}\left(\mathbf{m}-\mathbf{m_p}\right)^{\mathbf{T}} \mathbf{C_m^{-1}}(\boldsymbol{\xi})\left(\mathbf{m}-\mathbf{m_p}\right)\right], \tag{10}$$

where $\mathbb{P}[\boldsymbol{\xi}]$ encapsulates any preferences we wish to express regarding the covariance matrix. For convenience, and by convention, we refer to the elements of $\boldsymbol{\xi}$ as 'hyperparameters', and thus $\mathbb{P}[\boldsymbol{\xi}]$ is known as a 'hyperprior'.

In this expanded parameter space, the posterior distribution may be denoted $\mathbb{P}[\mathbf{m}, \boldsymbol{\xi} \mid \mathbf{d_0}]$. Marginalizing out the hyperparameters, and then applying the rules of conditional probability, we therefore have

$$\begin{aligned}\mathbb{P}[\mathbf{m} \mid \mathbf{d_0}] &= \int \mathbb{P}[\mathbf{m}, \boldsymbol{\xi} \mid \mathbf{d_0}]\; \mathrm{d}^K\boldsymbol{\xi} \\ &= \int \mathbb{P}[\mathbf{m} \mid \mathbf{d_0}, \boldsymbol{\xi}]\,\mathbb{P}[\boldsymbol{\xi} \mid \mathbf{d_0}]\; \mathrm{d}^K\boldsymbol{\xi}\,,\end{aligned} \tag{11}$$

with integration over all of $\boldsymbol{\xi}$-space. Of the quantities in this expression, the first has already been encountered: $\mathbb{P}[\mathbf{m} \mid \mathbf{d_0}, \boldsymbol{\xi}]$ is simply the posterior distribution given in eq. (9), but with the covariance matrix specified as $\mathbf{C_m}(\boldsymbol{\xi})$. However, the quantity $\mathbb{P}[\boldsymbol{\xi} \mid \mathbf{d_0}]$ is new, and deserves to be considered in more detail.

2.4.1 Identifying hyperparameters compatible with observations

To make sense of the hierarchical procedure, we must recognize that whenever we introduce a prior distribution in the model space, we are also implicitly defining a distribution in the data-space. If—for example—we choose to express a belief that the true system is somehow 'close' to a certain model, this implies that we expect observations to be similarly 'close' to the predictions that we can make for that model. For any given value of $\boldsymbol{\xi}$, the distribution $\mathbb{P}[\mathbf{m} \mid \boldsymbol{\xi}]$ represents a Gaussian in model-space, centred on $\mathbf{m_p}$ and with covariance $\mathbf{C_m}(\boldsymbol{\xi})$ (as in eq. 10). Any linear transformation of a Gaussian-distributed quantity yields another Gaussian-distributed quantity: if $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{Ax} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathbf{T}})$. The linear forward problem therefore maps our prior into a Gaussian in the data-space, and accounting for Gaussian data noise, we have

$$\mathbb{P}[\mathbf{d} \mid \boldsymbol{\xi}] = \frac{1}{\sqrt{(2\pi)^N\,|(\mathbf{G}\mathbf{C_m}(\boldsymbol{\xi})\mathbf{G}^{\mathbf{T}} + \mathbf{C_d})|}} \exp\left[-\frac{1}{2}\left(\mathbf{d}-\mathbf{G}\mathbf{m_p}\right)^{\mathbf{T}}\left(\mathbf{G}\mathbf{C_m}(\boldsymbol{\xi})\mathbf{G}^{\mathbf{T}} + \mathbf{C_d}\right)^{-\mathbf{1}}\left(\mathbf{d}-\mathbf{G}\mathbf{m_p}\right)\right]. \tag{12}$$

Thus, for any given choice of $\boldsymbol{\xi}$, one may make a quantitative assessment of whether the observed data, $\mathbf{d_0}$, is in some sense 'unexpected'. If $\mathbb{P}[\mathbf{m} \mid \boldsymbol{\xi}_1]$ and $\mathbb{P}[\mathbf{m} \mid \boldsymbol{\xi}_2]$ represent results based on two conflicting assessments of prior knowledge about a system, it may be appropriate to use $\mathbb{P}[\mathbf{d_0} \mid \boldsymbol{\xi}_1]$ and $\mathbb{P}[\mathbf{d_0} \mid \boldsymbol{\xi}_2]$ to help assess which is 'more reasonable'.

Within the hierarchical Bayesian procedure, we use this argument to assess the extent to which values of $\boldsymbol{\xi}$ are consistent with available evidence. Applying Bayes' theorem again, $\mathbb{P}[\boldsymbol{\xi} \mid \mathbf{d_0}]$ in eq. (11) can be rewritten

$$\mathbb{P}[\boldsymbol{\xi} \mid \mathbf{d_0}] = \frac{\mathbb{P}[\mathbf{d_0} \mid \boldsymbol{\xi}]\,\mathbb{P}[\boldsymbol{\xi}]}{\mathbb{P}[\mathbf{d_0}]}\,. \tag{13}$$

Thus, re-stating eq. (11) and summarizing the above discussion, the posterior distribution we seek may be expressed in the form

$$\mathbb{P}\left[\mathbf{m} \mid \mathbf{d_0}\right] = \int \mathbb{P}\left[\mathbf{m} \mid \mathbf{d_0}, \boldsymbol{\xi}\right] \mathbb{P}\left[\boldsymbol{\xi} \mid \mathbf{d_0}\right] \, \mathrm{d}^K \boldsymbol{\xi} \tag{14a}$$

$$\propto \int \mathbb{P}\left[\mathbf{m} \mid \mathbf{d_0}, \boldsymbol{\xi}\right] \mathbb{P}\left[\mathbf{d_0} \mid \boldsymbol{\xi}\right] \mathbb{P}\left[\boldsymbol{\xi}\right] \mathrm{d}^K \boldsymbol{\xi} \, , \tag{14b}$$

where $\mathbb{P}\left[\mathbf{m} \mid \mathbf{d_0}, \boldsymbol{\xi}\right]$ is as given in eq. (9), $\mathbb{P}\left[\mathbf{d_0} \mid \boldsymbol{\xi}\right]$ is as given in eq. (12), and where $\mathbb{P}\left[\boldsymbol{\xi}\right]$ is a hyperprior chosen to reflect any existing knowledge or preferences.

2.4.2 Obtaining approximate posterior distributions

A complete solution to the inverse problem therefore requires integration over the entire parameter space of $\boldsymbol{\xi}$, and in general this will not have an analytic solution. For simple problems, it may be possible to use a numerical integration technique—such as Monte Carlo integration—to perform the integral in eq. (14b) and directly characterize the posterior. Alternatively, it may be possible to use an approximation to $\mathbb{P}\left[\boldsymbol{\xi} \mid \mathbf{d_0}\right]$ that simplifies calculations. In particular, for covariance matrices parametrized by only a small number of parameters—such as the Tikhonov family—empirical evidence suggests that $\mathbb{P}\left[\boldsymbol{\xi} \mid \mathbf{d_0}\right]$ generally has a single, sharply defined peak, at some point which we denote by $\boldsymbol{\xi_0}$. Thus, one reasonable approximation may be a delta-distribution, $\mathbb{P}\left[\boldsymbol{\xi} \mid \mathbf{d_0}\right] \approx \delta(\boldsymbol{\xi} - \boldsymbol{\xi_0})$, in which case

$$\mathbb{P}\left[\mathbf{m} \mid \mathbf{d_0}\right] \approx \mathbb{P}\left[\mathbf{m} \mid \mathbf{d_0}, \boldsymbol{\xi_0}\right] \, , \tag{15}$$

and the solution can be obtained by using eq. (9) with covariance matrix $\mathbf{C_m}(\boldsymbol{\xi_0})$. Another reasonable approximation may be a Gaussian centred on $\boldsymbol{\xi_0}$,

$$\mathbb{P}\left[\boldsymbol{\xi} \mid \mathbf{d_0}\right] \approx \frac{1}{\sqrt{(2\pi)^K \left|\boldsymbol{\Phi}\right|}} \exp\left[-\frac{1}{2}\left(\boldsymbol{\xi} - \boldsymbol{\xi_0}\right)^{\mathbf{T}} \boldsymbol{\Phi}^{-\mathbf{1}} \left(\boldsymbol{\xi} - \boldsymbol{\xi_0}\right)\right] , \tag{16a}$$

with the covariance matrix Φ chosen based on the curvature of $\mathbb{P}\left[\boldsymbol{\xi} \mid \mathbf{d_0}\right]$ at the peak,

$$\left[\boldsymbol{\Phi}^{-\mathbf{1}}\right]_{ij} = - \left. \frac{\partial^2 \log \mathbb{P}\left[\boldsymbol{\xi} \mid \mathbf{d_0}\right]}{\partial \xi_i \partial \xi_j} \right|_{\boldsymbol{\xi} = \boldsymbol{\xi_0}} . \tag{16b}$$

If this Gaussian approximation is used, eq. (14a) may be relatively straightforward to evaluate using a Gauss–Hermite quadrature rule (e.g Abramowitz & Stegun 1964).

3 IMPLEMENTATION

In order to implement the hierarchical procedure efficiently, we need to be able to determine $\boldsymbol{\xi_0}$, the location of the maximum of the distribution $\mathbb{P}\left[\boldsymbol{\xi} \mid \mathbf{d_0}\right]$. In this section, we derive the necessary expressions—first for a general $\mathbf{C_m}(\boldsymbol{\xi})$, and then for special cases of common interest. A number of additional formulae that are necessary for computations, but which are of little importance to our discussion, have been relegated to Appendix A. Readers who wish to follow the detailed derivation of these results may benefit from the compendium of matrix identities (including results from matrix calculus) compiled by Petersen & Pedersen (2015): we use various of these results. In this section, we focus only on deriving the necessary formulae; discussion of their interpretation is postponed until Section 5.1.

3.1 The general case

We begin by considering the general case, $\mathbf{C_m} = \mathbf{C_m}(\boldsymbol{\xi})$. It can be shown that $\mathbb{P}\left[\mathbf{d_0} \mid \boldsymbol{\xi}\right]$ (as in eq. 12) is equivalent to

$$\mathbb{P}\left[\mathbf{d}' \mid \mathbf{C_m^{-1}}\right] = \sqrt{\frac{\left|\mathbf{C_m^{-1}}\right| \left|\mathbf{C_d^{-1}}\right|}{(2\pi)^N \left|\mathbf{G^T C_d^{-1} G} + \mathbf{C_m^{-1}}\right|}} \exp\left[-\frac{1}{2}\mathbf{d}'^{\mathbf{T}} \mathbf{C_d^{-1}} \mathbf{d}'\right] \exp\left[\frac{1}{2}\mathbf{d}'^{\mathbf{T}} \mathbf{C_d^{-1}} \mathbf{G} \left(\mathbf{G^T C_d^{-1} G} + \mathbf{C_m^{-1}}\right)^{-1} \mathbf{G^T C_d^{-1}} \mathbf{d}'\right] , \tag{17}$$

where again $\mathbf{d}' = \mathbf{d_0} - \mathbf{G m_p}$, and where the dependence of $\mathbf{C_m}$ on $\boldsymbol{\xi}$ is omitted for brevity. We note that all the quantities contained within this expression are routinely available within any computational implementation of the least-squares algorithm. It will be convenient to work with the (natural) logarithm of this probability density,

$$\begin{aligned} \log \mathbb{P}\left[\mathbf{d}' \mid \mathbf{C_m^{-1}}\right] = \frac{1}{2} \Big\{ \Big(\log \left|\mathbf{C_d^{-1}}\right| - N \log 2\pi - \mathbf{d}'^{\mathbf{T}} \mathbf{C_d^{-1}} \mathbf{d}' \Big) + \log \left|\mathbf{C_m^{-1}}\right| - \log \left|\mathbf{G^T C_d^{-1} G} + \mathbf{C_m^{-1}}\right| \\ + \mathbf{d}'^{\mathbf{T}} \mathbf{C_d^{-1}} \mathbf{G} \left(\mathbf{G^T C_d^{-1} G} + \mathbf{C_m^{-1}}\right)^{-1} \mathbf{G^T C_d^{-1}} \mathbf{d}' \Big\} , \end{aligned} \tag{18}$$

where the parenthesized terms at the start of this expression are independent of $\mathbf{C_m}$. Differentiating with respect to the individual elements of $\mathbf{C_m^{-1}}$, we obtain

$$\frac{\partial \log \mathbb{P}\left[\mathbf{d}' \mid \mathbf{C_m^{-1}}\right]}{\partial \mathbf{C_m^{-1}}} = \frac{1}{2} \left\{ \mathbf{C_m} - \left(\mathbf{G^T C_d^{-1} G} + \mathbf{C_m^{-1}}\right)^{-1} - \left(\mathbf{G^T C_d^{-1} G} + \mathbf{C_m^{-1}}\right)^{-1} \mathbf{G^T C_d^{-1}} \mathbf{d}' \mathbf{d}'^{\mathbf{T}} \mathbf{C_d^{-1}} \mathbf{G} \left(\mathbf{G^T C_d^{-1} G} + \mathbf{C_m^{-1}}\right)^{-1} \right\} , \tag{19}$$

where the notation $\mathbf{Y} = \partial f / \partial \mathbf{X}$ should be understood to imply that $Y_{ij} = \partial f / \partial X_{ij}$. Then, we can apply the chain rule for matrix differentiation to obtain

$$\frac{\partial \log \mathbb{P}\left[\mathbf{d}' \mid \boldsymbol{\xi}\right]}{\partial \xi_i} = \mathrm{Tr}\left\{\left[\frac{\partial \mathbf{C}_\mathbf{m}^{-1}}{\partial \xi_i}\right]^\mathrm{T} \frac{\partial \log \mathbb{P}\left[\mathbf{d}' \mid \mathbf{C}_\mathbf{m}^{-1}\right]}{\partial \mathbf{C}_\mathbf{m}^{-1}}\right\}. \tag{20}$$

We highlight the fact that it is the inverse of $\mathbf{C_m}$, rather than $\mathbf{C_m}$ itself, that appears in this expression. Finally, from eq. (13) we have

$$\frac{\partial \log \mathbb{P}\left[\boldsymbol{\xi} \mid \mathbf{d_0}\right]}{\partial \xi_i} = \mathrm{Tr}\left\{\left[\frac{\partial \mathbf{C}_\mathbf{m}^{-1}}{\partial \xi_i}\right]^\mathrm{T} \frac{\partial \log \mathbb{P}\left[\mathbf{d}' \mid \mathbf{C}_\mathbf{m}^{-1}\right]}{\partial \mathbf{C}_\mathbf{m}^{-1}}\right\} + \frac{\partial \log \mathbb{P}\left[\boldsymbol{\xi}\right]}{\partial \xi_i}. \tag{21}$$

Armed with eq. (18) and these derivatives, it is conceptually straightforward to use a numerical optimization algorithm to maximize $\mathbb{P}\left[\mathbf{d_0} \mid \boldsymbol{\xi}\right]$. Of course, whether this is computationally tractable depends on a number of factors, especially the number of hyperparameters K, and the model-space dimension M. As we demonstrate below, more efficient methods of solution may be possible in specific cases.

3.2 Tikhonov-style regularization

A common choice for $\mathbf{C_m}$ will be the Tikhonov regularization family. We therefore derive specialized results for this case, which corresponds to choosing

$$\mathbf{C}_\mathbf{m}^{-1}(\alpha, \beta) = \alpha^2 \mathbf{I} + \beta^2 \mathbf{H}, \tag{22}$$

as in eq. (6). Applying eq. (20), it is straightforward to determine that

$$\frac{\partial \log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right]}{\partial \alpha} = \alpha\, \mathrm{Tr}\left\{\left(\alpha^2 \mathbf{I} + \beta^2 \mathbf{H}\right)^{-1} - \left(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathbf{d}^{-1}\mathbf{G} + \alpha^2 \mathbf{I} + \beta^2 \mathbf{H}\right)^{-1}\right\} - \alpha \mathbf{d}'^\mathrm{T}\mathbf{C}_\mathbf{d}^{-1}\mathbf{G}\left(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathbf{d}^{-1}\mathbf{G} + \alpha^2 \mathbf{I} + \beta^2 \mathbf{H}\right)^{-2} \mathbf{G}^\mathrm{T}\mathbf{C}_\mathbf{d}^{-1}\mathbf{d}', \tag{23a}$$

where we have introduced the notation $\mathbf{M^{-2}} = \mathbf{M^{-1}}\mathbf{M^{-1}}$, and made use of the cyclic invariance property of the trace. Similarly,

$$\begin{aligned}\frac{\partial \log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right]}{\partial \beta} &= \beta\, \mathrm{Tr}\left\{\mathbf{H}\left[\left(\alpha^2 \mathbf{I} + \beta^2 \mathbf{H}\right)^{-1} - \left(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathbf{d}^{-1}\mathbf{G} + \alpha^2 \mathbf{I} + \beta^2 \mathbf{H}\right)^{-1}\right]\right\} \\ &\quad - \beta \mathbf{d}'^\mathrm{T}\mathbf{C}_\mathbf{d}^{-1}\mathbf{G}\left(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathbf{d}^{-1}\mathbf{G} + \alpha^2 \mathbf{I} + \beta^2 \mathbf{H}\right)^{-1} \mathbf{H}\left(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathbf{d}^{-1}\mathbf{G} + \alpha^2 \mathbf{I} + \beta^2 \mathbf{H}\right)^{-1} \mathbf{G}^\mathrm{T}\mathbf{C}_\mathbf{d}^{-1}\mathbf{d}',\end{aligned} \tag{23b}$$

where we have made use of the fact that that $\mathbf{H}$ must be symmetric for $\mathbf{C}_\mathbf{m}^{-1}$ to represent a valid covariance matrix. As before, it is apparent that

$$\frac{\partial \log \mathbb{P}\left[\alpha, \beta \mid \mathbf{d_0}\right]}{\partial \alpha} = \frac{\partial \log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right]}{\partial \alpha} + \frac{\partial \log \mathbb{P}\left[\alpha, \beta\right]}{\partial \alpha}, \tag{24}$$

and similarly for differentiation with respect to β; these expressions will equal zero when $\mathbb{P}\left[\alpha, \beta \mid \mathbf{d}'\right]$ is maximized. To characterize the spread of $\mathbb{P}\left[\alpha, \beta \mid \mathbf{d_0}\right]$, it may be useful to compute the second partial derivatives of the distribution at the optimum: necessary expressions are stated in Appendix A.

3.2.1 Special case: fixed β

Although it is possible to implement optimization using eqs (18) and (23), this will entail an iterative procedure with matrix inversion at every step. For modest-scale inverse problems, this is, nevertheless, tractable; pre-conditioned solvers may prove beneficial as the model-space dimension increases. However, in the special case where β is fixed, we can obtain efficient algorithms by working in a spectral domain. This may be of particular interest in the case where a smoothing term is not desired, $\beta = 0$, although we consider the more general situation here.

We introduce two eigen-decompositions: $\mathbf{S}\boldsymbol{\Lambda}\mathbf{S}^\mathrm{T} = \mathbf{G}^\mathrm{T}\mathbf{C}_\mathbf{d}^{-1}\mathbf{G} + \beta^2\mathbf{H}$, and $\mathbf{T}\boldsymbol{\Omega}\mathbf{T}^\mathrm{T} = \mathbf{H}$, with $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1 \ldots, \lambda_M)$ and $\boldsymbol{\Omega} = \mathrm{diag}(\omega_1, \ldots, \omega_M)$. eq. (23a) then becomes

$$\begin{aligned}\frac{\partial \log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right]}{\partial \alpha} &= \alpha\, \mathrm{Tr}\left\{\mathbf{T} \cdot \mathrm{diag}\left(\frac{1}{\alpha^2 + \beta^2\omega_1}, \ldots, \frac{1}{\alpha^2 + \beta^2\omega_M}\right) \cdot \mathbf{T}^\mathrm{T} - \mathbf{S} \cdot \left(\frac{1}{\lambda_1 + \alpha^2}, \ldots, \frac{1}{\lambda_M + \alpha^2}\right) \cdot \mathbf{S}^\mathrm{T}\right\} \\ &\quad - \alpha \mathbf{d}'^\mathrm{T}\mathbf{C}_\mathbf{d}^{-1}\mathbf{G}\mathbf{S} \cdot \left(\frac{1}{(\lambda_1 + \alpha^2)^2}, \ldots, \frac{1}{(\lambda_M + \alpha^2)^2}\right) \cdot \mathbf{S}^\mathrm{T}\mathbf{G}^\mathrm{T}\mathbf{C}_\mathbf{d}^{-1}\mathbf{d}'.\end{aligned} \tag{25}$$

Since β is fixed, the partial derivative with respect to β does not arise. Using α_0 to denote the most-probable value of α, we use the requirement that $\partial \log \mathbb{P}\left[\alpha \mid \beta, \mathbf{d}'\right] / \partial \alpha\big|_{\alpha_0} = 0$ to obtain

$$\sum_{i=1}^{M} \alpha_0 \left\{\frac{1}{\alpha_0^2 + \beta^2\omega_i} - \frac{1}{\lambda_i + \alpha_0^2} - \frac{\left[\mathbf{S}^\mathrm{T}\mathbf{G}^\mathrm{T}\mathbf{C}_\mathbf{d}^{-1}\mathbf{d}'\right]_i^2}{(\lambda_i + \alpha_0^2)^2}\right\} + \left.\frac{\partial \log \mathbb{P}\left[\alpha \mid \beta\right]}{\partial \alpha}\right|_{\alpha_0} = 0. \tag{26}$$

This is straightforwardly—and cheaply—solved using a numerical root-finding algorithm. Note that if $\beta = 0$, the first term in this sum must still be included M times, to give a total contribution M/α_0^2. Expressions for $\log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right]$ and the second derivative $\partial^2 \log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right] / \partial \alpha^2$ in terms of this eigendecomposition are recorded in Appendix A2.

3.2.2 Special case: fixed α

Although perhaps less frequently encountered, we can perform a similar analysis for the case where α is fixed, possibly to zero. Provided $\mathbf{H}$ is invertible, we can introduce a further eigendecomposition, $\mathbf{U\Gamma U^{-1}} = \mathbf{H^{-1}}\left(\mathbf{G^T C_d^{-1} G} + \alpha^2\mathbf{I}\right)$, with $\mathbf{\Gamma} = \mathrm{diag}(\gamma_1, \ldots, \gamma_M)$. We note that this matrix is non-symmetric, and therefore its eigenvectors need not be orthogonal. We can then write $\left(\mathbf{G^T C_d^{-1} G} + \alpha^2\mathbf{I} + \beta^2\mathbf{H}\right) = \mathbf{HU}\left(\mathbf{\Gamma} + \beta^2\mathbf{I}\right)\mathbf{U^{-1}}$. Thus, eq. (23b) can be expressed

$$\frac{\partial \log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right]}{\partial \beta} = \beta\,\mathrm{Tr}\left\{\mathbf{T}\cdot\mathrm{diag}\left(\frac{\omega_1}{\alpha^2+\beta^2\omega_1}, \ldots, \frac{\omega_M}{\alpha^2+\beta^2\omega_M}\right)\cdot\mathbf{T^T} - \mathbf{HU}\cdot\mathrm{diag}\left(\frac{1}{\gamma_1+\beta^2}, \ldots, \frac{1}{\gamma_M+\beta^2}\right)\cdot\mathbf{U^{-1}H^{-1}}\right\}$$
$$- \beta\mathbf{d'}^{\mathbf{T}}\mathbf{C_d^{-1}GU}\cdot\mathrm{diag}\left(\frac{1}{(\gamma_1+\beta^2)^2}, \ldots, \frac{1}{(\gamma_M+\beta^2)}\right)\cdot\mathbf{U^{-1}H^{-1}G^T C_d^{-1}d'}\,. \tag{27}$$

Again, we use β_0 to represent an optimal value of β, which is the solution to another root-finding problem,

$$\sum_{i=1}^{M}\beta_0\left\{\frac{\omega_i}{\alpha^2+\beta_0^2\omega_i} - \frac{1}{\gamma_i+\beta_0^2} - \frac{\left[\mathbf{U^T G^T C_d^{-1} d'}\right]_i\left[\mathbf{U^{-1}H^{-1}G^T C_d^{-1} d'}\right]_i}{\left(\gamma_i+\beta_0^2\right)^2}\right\} + \left.\frac{\partial \log \mathbb{P}\left[\beta \mid \alpha\right]}{\partial\beta}\right|_{\beta_0} = 0\,. \tag{28}$$

Expressions for $\log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right]$ and the second derivative $\partial^2 \log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right]/\partial\beta^2$ can be found in Appendix A3.

3.3 Automatic relevance detetermination

Another specific case of interest is what Mackay (1992a) terms 'automatic relevance determination'. This amounts to using a diagonal covariance matrix, with each model parameter (or group of parameters) being assigned its own uncertainty parameter,

$$\mathbf{C_m}(\boldsymbol{\xi}) = \mathrm{diag}(\xi_1, \xi_2, \ldots \xi_M)\,. \tag{29}$$

By adopting a hyperprior on $\boldsymbol{\xi}$ that expresses a preference for the elements of $\boldsymbol{\xi}$ to remain close to zero (and thus, for the elements of the inverse covariance matrix to be large), maximizing $\mathbb{P}\left[\boldsymbol{\xi} \mid \mathbf{d_0}\right]$ causes model parameters that are unnecessary to explain the data to be 'regularized out' of the solution. In effect, this provides a mechanism for achieving sparse solutions to the inverse problem within an L_2 minimization framework.

To obtain the necessary expressions for optimization, it is convenient to note that

$$\left[\frac{\partial \mathbf{C_m^{-1}}}{\partial \xi_k}\right]_{ij} = \frac{-\delta_{ij}\delta_{ik}}{\xi_i^2}\,. \tag{30}$$

Thus,

$$\frac{\partial \log \mathbb{P}\left[\boldsymbol{\xi} \mid \mathbf{d_0}\right]}{\partial \xi_i} = \frac{1}{2\xi_i}\left\{\frac{1}{\xi_i}\left\{\left[\left(\mathbf{G^T C_d^{-1} G} + \mathbf{C_m^{-1}}\right)^{-1}\right]_{ii} + \left[\left(\mathbf{G^T C_d^{-1} G} + \mathbf{C_m^{-1}}\right)^{-1}\mathbf{G^T C_d^{-1} d'}\right]_i^2\right\} - 1\right\} + \frac{\partial \log \mathbb{P}\left[\boldsymbol{\xi}\right]}{\partial \xi_i}\,, \tag{31}$$

where no summation is implied by the repetition of an index. A more general case is found by allowing the model to be sparse in a different basis from that used for inversion. This corresponds to choosing

$$\mathbf{C_m}(\boldsymbol{\xi}) = \mathbf{J}\cdot\mathrm{diag}(\xi_1, \xi_2, \ldots, \xi_K)\cdot\mathbf{J^T}\,, \tag{32}$$

where $\mathbf{J}$ is an $M \times K$ matrix satisfying the property $\mathbf{J^T J} = \mathbf{I}_K$, the K-dimensional identity matrix, so that

$$\mathbf{C_m^{-1}}(\boldsymbol{\xi}) = \mathbf{J}\cdot\mathrm{diag}\left(\frac{1}{\xi_1}, \frac{1}{\xi_2}\ldots\frac{1}{\xi_K}\right)\cdot\mathbf{J^T}\,. \tag{33}$$

This results in a similar expression for the partial derivative,

$$\frac{\partial \log \mathbb{P}\left[\boldsymbol{\xi} \mid \mathbf{d_0}\right]}{\partial \xi_i} = \frac{1}{2\xi_i}\left\{\frac{1}{\xi_i}\left\{\left[\mathbf{J^T}\left(\mathbf{G^T C_d^{-1} G} + \mathbf{C_m^{-1}}\right)^{-1}\mathbf{J}\right]_{ii} + \left[\mathbf{J^T}\left(\mathbf{G^T C_d^{-1} G} + \mathbf{C_m^{-1}}\right)^{-1}\mathbf{G^T C_d^{-1} d'}\right]_i^2\right\} - 1\right\} + \frac{\partial \log \mathbb{P}\left[\boldsymbol{\xi}\right]}{\partial \xi_i}\,. \tag{34}$$

We remark that this can be simplified further in the event that $\mathbf{J}$ is chosen to correspond to the eigenvectors of $\mathbf{G^T C_d^{-1} G}$, although it is not obvious that this is generally an appropriate structure to impose upon the covariance matrix.

4 DEMONSTRATION

To demonstrate the preceding theory effectively, it is convenient to choose an inverse problem that is amenable to illustration in two dimensions. Unfortunately, few geophysically meaningful problems exhibit such simplicity, and we therefore consider a simple polynomial regression problem. We generate noisy random samples from a known function, and then perform an inversion to recover the coefficients of a polynomial that best explains this data set. Specifically, we generate $N = 10$ x-values uniformly at random in the range $(-1, 1)$, and then compute $f(x) = \frac{1}{4}x^7 + 2x^4 - x + 1$ for each—as shown in Figs 1(a) and (b). We add uncorrelated zero-mean Gaussian random noise with standard

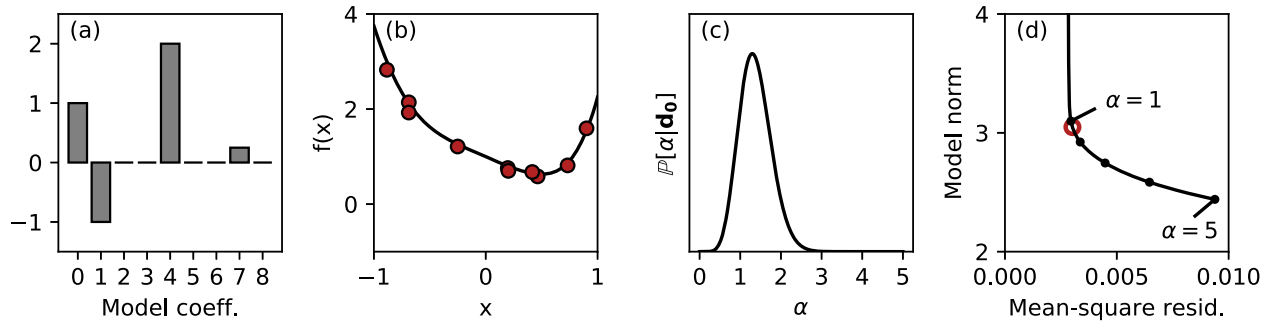

Figure 1. A simple regularized inverse problem illustrating the distribution $\mathbb{P}[\alpha \mid \mathbf{d_0}]$. (a) The function $f(x) = \frac{1}{4}x^7 + 2x^4 - x + 1$ is represented by a model vector of polynomial coefficients. (b) Noisy samples are generated from this function, using an uncorrelated Gaussian noise model with standard deviation $\sigma = 0.1$. (c) Assuming $\mathbf{C_m^{-1}} = \alpha^2\mathbf{I}$, we plot the distribution $\mathbb{P}[\alpha \mid \mathbf{d_0}]$ for a uniform hyperprior on α, obtained using eq. (17). (d) We can also plot the tradeoff curve between model norm and model misfit; black dots denote $\alpha = 1 \ldots 5$, and the red circle corresponds to the peak of $\mathbb{P}[\alpha \mid \mathbf{d_0}]$. The distribution has a well-defined peak, which corresponds well to the 'elbow' of the L-curve.

deviation $\sigma = 0.1$ to each value for $f(x)$. This forms a data set for inversion, with the inverse problem parametrized as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^P \\ 1 & x_2 & x_2^2 & \cdots & x_2^P \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^P \end{pmatrix} \begin{pmatrix} m_0 \\ m_1 \\ m_2 \\ \vdots \\ m_P \end{pmatrix}. \tag{35}$$

Note that we have indexed the model coefficients starting at 0, so that $y(x) = \sum_n m_n x^n$, and thus, the dimension of the model space is $M = P + 1$; our inversion assumes $P = 8$.

4.1 Tikhonov-style regularization

We assume $\mathbf{C_d} = \sigma^2\mathbf{I}$—in other words, the noise model assumed for inversion matches the noise actually present in the data. The unregularized matrix $\mathbf{G^TG}$ for this problem turns out to be ill-conditioned, with condition number (ratio of largest- to smallest eigenvalue) $\kappa = 5.4 \times 10^{10}$. Thus, regularization is necessary to obtain a sensible solution, and we use 'ridge regression', $\mathbf{C_m^{-1}} = \alpha^2\mathbf{I}$, with $\mathbf{m_p} = \mathbf{0}$; a smoothing term is not appropriate for this problem. Using eq. (17), it is then straightforward to map out the distribution $\mathbb{P}[\alpha \mid \mathbf{d_0}]$ for this example, adopting a uniform, improper hyperprior on α. The result is shown in Fig. 1(c), and we see that this has a single, well-defined peak at $\alpha_0 \approx 1.3$, with plausible values roughly in the range $0.5 < \alpha < 2.5$. Since the model-space has low dimension, it is also computationally cheap to map out the progression of recovered model norm ($\mathbf{m^Tm}$) and mean-squared residuals ($(\mathbf{d} - \mathbf{Gm})^\mathbf{T}(\mathbf{d} - \mathbf{Gm})/N$) as α is varied. Plotting one quantity against the other results in an L-curve (Fig. 1d) with an 'elbow' consistent with this range. Thus, the hierarchical Bayesian procedure appears to indicate regularization parameters consistent with those that might be obtained via existing methods—but our approach is arguably more rigorously defined.

Applying eq. (14), $\mathbb{P}[\mathbf{m} \mid \mathbf{d}]$ can be obtained by performing an integral over α. Marginal distributions $\mathbb{P}[m_i \mid \mathbf{d_0}] = \int\cdots\int \mathbb{P}[\mathbf{m} \mid \mathbf{d_0}] \prod_{j \neq i} dm_j$ are shown in black in Fig. 2, along with the true value of each parameter (as a vertical line). Alternatively, as described in Section 2.4.2, the distribution $\mathbb{P}[\alpha \mid \mathbf{d_0}]$ (e.g. Fig. 1c) can be approximated by a δ-distribution. In practical terms, this is equivalent to selecting the most-probable value of α and using this to perform an inversion according to eq. (9). The resulting approximate posterior marginal distributions are shown in blue in Fig. 2, and are almost indistinguishable from the 'complete' marginals. This is an attractive result, as the integration over hyperparameters required to evaluate eq. (14) can be computationally challenging.

4.1.1 Increased data noise

We now repeat this experiment, except with the noise level increased by an order of magnitude (i.e. $\sigma = 1$). Again, we assume this same noise level during inversion, $\mathbf{C_d^{-1}} = \sigma^2\mathbf{I}$, and a counterpart to Fig. 1 is shown in Fig. 3. The distribution $\mathbb{P}[\alpha \mid \mathbf{d_0}]$ continues to have a well-defined peak (now at $\alpha_0 \approx 1.5$), and agrees well with choices that might be made based on the L-curve. However, the distribution is less tightly constrained, and assigns appreciable probability mass to even large values of α. This arises because—with a high enough noise level—the data could plausibly be generated from the prior mean model $\mathbf{m_p} = \mathbf{0}$. Changing the width of the prior does not alter the location of this *a priori* most probable model, and hence the data provides only relatively weak constraint upon α.

This introduces difficulties for the computation of $\mathbb{P}[\mathbf{m} \mid \mathbf{d_0}]$, and we need to provide a more rigorous specification of the hyperprior, $\mathbb{P}[\alpha]$. In the preceding example, we effectively adopted an improper prior: any positive value of α was regarded as equally likely. Now, we must introduce an upper bound on α to ensure that the integral in eq. (14) converges. For the sake of argument, we assume that $\mathbb{P}[\alpha] = 0$ for

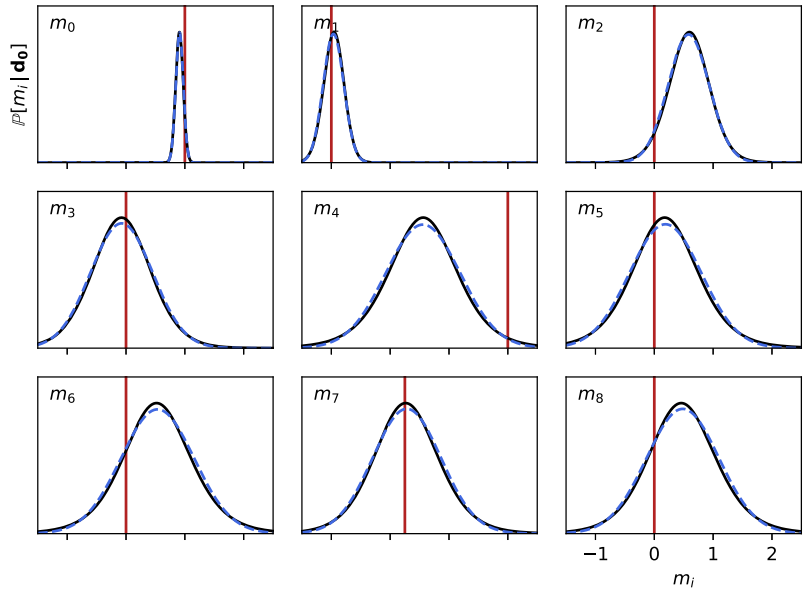

Figure 2. Recovery of model parameters. Marginal distributions for each model parameter separately, for inverse problem described in Fig. 1. Distributions depicted in black are exact representations of $\mathbb{P}[\mathbf{m} \mid \mathbf{d_0}]$, obtained by integrating over all possible values of α as in eq. (14). Dashed, blue distributions are obtained by using a δ-distribution approximation to $\mathbb{P}[\alpha \mid \mathbf{d_0}]$ (as in eq. 15); the two are almost indistinguishable. Red vertical bars indicate the 'true' value of each model parameter. Each subplot is scaled to display the same maximum amplitude.

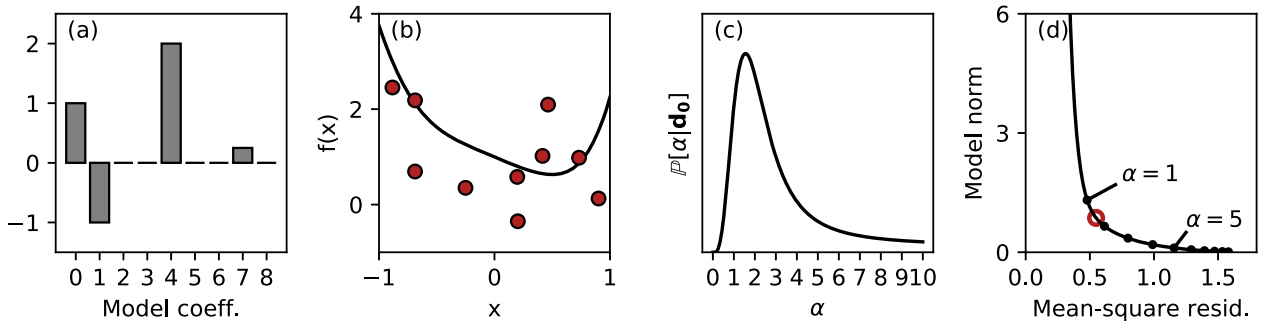

Figure 3. Inversion with higher noise levels. (a, b) As Fig. 1, except that data noise is generated from a distribution with $\sigma = 1$. (c) The distribution $\mathbb{P}[\alpha \mid \mathbf{d_0}]$ continues to have a well-defined peak, but now has a heavy tail; note that the α-range shown here is twice that in Fig. 1. (d) Nevertheless, the location of the peak corresponds well to the 'elbow' of the L-curve (dots denote $\alpha = 1, \ldots, 10$).

$\alpha > 10$, with uniform probability in the range $0 < \alpha < 10$ as illustrated in Fig. 3(d). Under this assumption, we can again compute posterior marginal distributions for each model parameter, shown in black in Fig. 4. These are notably non-Gaussian in form, with high probability mass assigned close to the prior model, $\mathbf{m_p}$. They are also substantially different from the (blue) approximate marginals obtained using eq. (9) and the maximum-likelihood value of α. Nevertheless, the maxima and spreads of the two sets of distributions are broadly similar, and the approximate versions are likely to be 'good enough' for many practical purposes. We remark that the differences increase if the prior on α is broadened, so that more high-α examples are included.

4.2 Automatic relevance determination

Returning to the original noise level ($\sigma = 0.1$, as in Fig. 1), we now adopt a different regularization strategy—that of automatic relevance determination, as introduced in Section 3.3. To allow straightforward comparison to the preceding experiments, where the hyperparameters were used to directly specify the inverse model covariance matrix, we choose $\mathbf{C}_\mathbf{m}^{-1}(\boldsymbol{\theta}) = \mathrm{diag}(\theta_0^2, \theta_1^2, \ldots, \theta_P^2)$. This is a slightly different parametrization from that analysed earlier: in effect, we use $\xi_i = 1/\theta_i^2$. The distribution $\mathbb{P}[\boldsymbol{\theta} \mid \mathbf{d_0}]$ is defined in a 9-D space, and cannot be plotted directly. However, it is straightforward to determine the maximum-likelihood value, and then plot the 1-D conditional distributions $\mathbb{P}\left[\theta_i \mid \mathbf{d_0}, \{\theta_j = \theta_j^0, j \neq i\}\right]$ which can be viewed as 'cross-sections' through the full distribution at the maximum-likelihood point. These are shown in Fig. 5, and can be separated into two classes: those where a clear maximum can be identified for relatively small values of θ_i, and those where the distribution is flat or growing as θ_i increases. We note that a preference for a large value of θ_i for a given model parameter implies a preference for fixing that parameter to its *a priori* value.

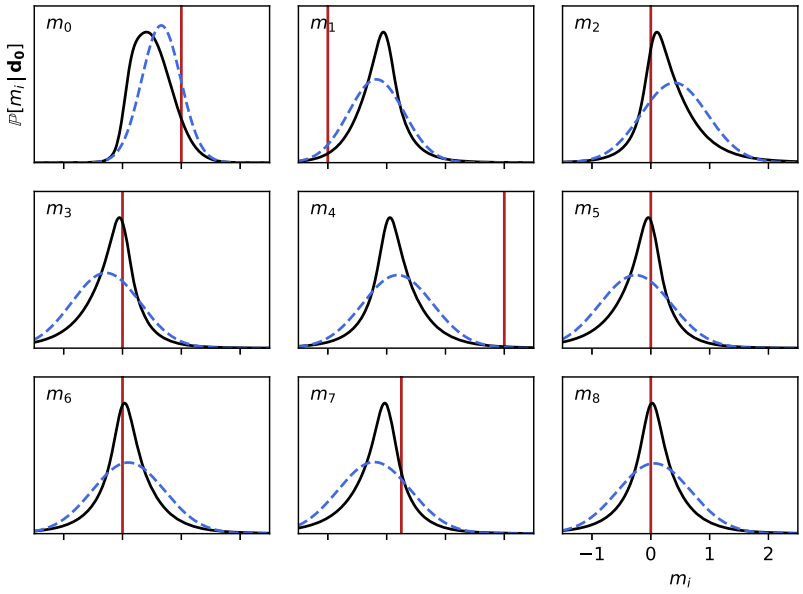

Figure 4. Posterior marginal distributions obtained under higher noise levels; compare Fig. 2. Black distributions represent exact versions of $\mathbb{P}[\mathbf{m} \mid \mathbf{d_0}]$; dashed blue distributions are approximations obtained by replacing $\mathbb{P}[\alpha \mid \mathbf{d_0}]$ by a δ-distribution centred on the maximum-likelihood regularizer. Red vertical bars represent the 'true' solution to this synthetic inverse problem.

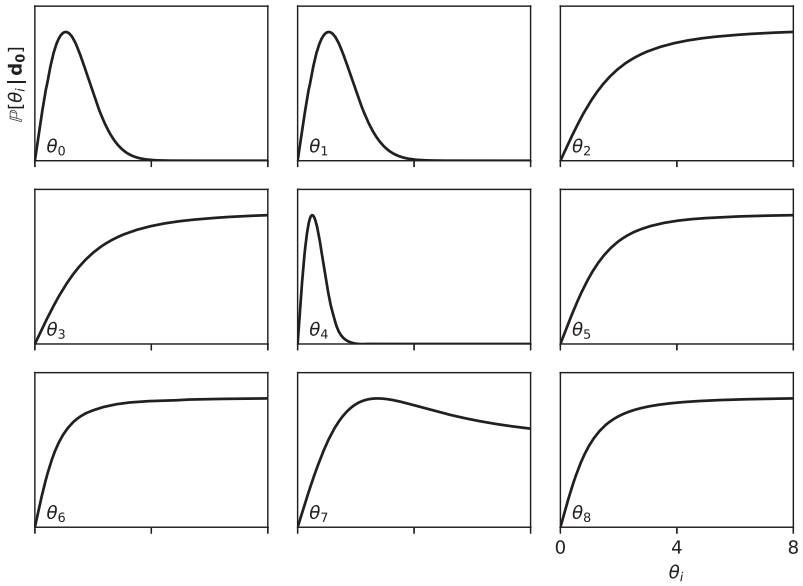

Figure 5. Automatic relevance determination: the distribution $\mathbb{P}[\boldsymbol{\theta} \mid \mathbf{d_0}]$. Each panel depicts a 1-D 'slice' through the distribution at the maximum-likelihood point (i.e. the conditional distribution $\mathbb{P}\left[\theta_i \mid \mathbf{d_0}, \{\theta_j = \theta_j^0, j \neq i\}\right]$). A uniform, improper prior on $\boldsymbol{\theta}$ is assumed. For parameters m_0, m_1, m_4 and m_7 a finite maximum can be identified. Maxima of the other parameters are effectively 'at infinity'. This implies that the most-probable regularizer will fix these components of the inversion to their *a priori* values.

Given the relative complexity of $\mathbb{P}[\boldsymbol{\theta} \mid \mathbf{d_0}]$, it is not straightforward to perform the integration required to compute $\mathbb{P}[\mathbf{m} \mid \mathbf{d_0}]$ according to eq. (14), and we do not attempt to do so. However, it is easy to compute posterior marginal distributions by adopting the maximum-likelihood regularizer and applying eq. (9). These are shown in Fig. 6, and exhibit remarkable accuracy: the inversion is able to identify the four non-zero components of the model, while constraining other parameters to zero. We remark that the 'true' posterior distributions are probably somewhat broader than those obtained using the approximation: such an effect would be consistent with what was seen in Figs 2 and 4. Results using Automatic Relevance Determination for noisier data are somewhat less impressive, perhaps because—as in Fig. 4—the δ-distribution approximation is less effective in this case.

Intriguingly, the Automatic Relevance Determination procedure is able to produce a sparse model—that is, one where many of the coefficients are zero. This is a property that can be physically desirable under certain circumstances, where processes are expected to be somehow localized. Sparsity-promoting inversion has attracted considerable attention in recent years, and underpins innovative concepts such as 'compressive sensing', where band-limited signals can be recovered from samples far below the Nyquist limit (e.g. Candès & Wakin

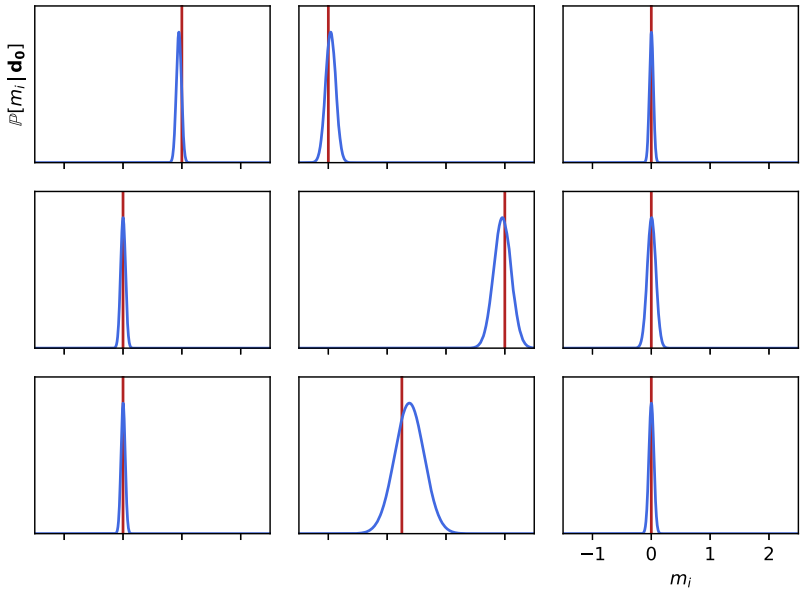

Figure 6. Approximate posterior marginals obtained using Automatic Relevance Determination. Each panel shows a 1-D marginal distribution, $\mathbb{P}[m_i \mid \mathbf{d_0}]$, obtained using a δ-distribution approximation to $\mathbb{P}[\boldsymbol{\theta} \mid \mathbf{d_0}]$ centred upon the maximum-likelihood value of $\boldsymbol{\theta}$. Red vertical lines correspond to 'true' values of model parameters. Almost perfect model recovery is achieved; contrast with Fig. 2.

2008). To properly quantify sparsity, one adopts an L_0 norm, which counts the number of non-zero components. However, this is not suited to use within optimization, and most recent progress has been built upon minimization of an L_1 norm: it has been shown that this is a very good approximation to minimization of the L_0 norm (Donoho 2006a,b). However, this necessitates the use of convex optimization procedures, which are computationally challenging. The Automatic Relevance Determination approach appears to provide a route towards obtaining sparse models within an L_2 minimization framework, which is significantly more straightforward to implement.

4.3 A geophysical example: dynamic topography

As a final example, we consider the problem of estimating the power spectrum of dynamic topography—the component of Earth's surface elevation which arises in response to geodynamic processes, such as uplift above a mantle upwelling. This was recently considered by Hoggard *et al.* (2016), who assembled a data set of point estimates and then performed a least-squares inversion to find the degree-30 spherical harmonic model that best explained this data set. The result showed substantially less power at low degree than was anticipated based on simulation results, and higher power at high degrees. Since Hoggard used Tikhonov regularization in the inversion, some debate has ensued about whether the discrepancy is simply due to the damping parameters chosen.

We repeat Hoggard's analysis, using his published data set, but determining Tikhonov regularization parameters using the techniques described within this paper. Using $Y_{lm}(\theta, \phi)$ to denote a real surface spherical harmonic, defined as in Hoggard *et al.* (2016), we treat each data point as arising from a spherical harmonic expansion,

$$d_i = \sum_{l=1}^{L} \sum_{m=-l}^{l} c_{lm} Y_{lm}(\theta_i, \phi_i), \tag{36}$$

where (θ_i, ϕ_i) denote the spatial coordinates corresponding to datum d_i. Choosing $L = 30$, and noting that an $l = 0$ term is not included in the expansion, we seek the $L(L + 2) = 960$ coefficients c_{lm}. For our prior, we follow Hoggard's approach, and adopt $\mathbf{C_m^{-1}} = \alpha^2\mathbf{I} + \beta^2\mathbf{H}$. The elements of $\mathbf{H}$ are given by $H_{ij} = \delta_{ij} l_i(l_i + 1)$, where l_i is the angular order associated with the ith model coefficient: this choice is derived from a requirement that the average gradient of the spherical harmonic expansion should not be allowed to become too large. Hoggard explored a range of regularization parameters, guided by L-curve analysis, and ultimately determined $\alpha = 20$, $\beta = 1$ to be a reasonable choice (although he reports results across a range of plausible regularizations). Now, we repeat his analysis, treating both as hyperparameters following the hierarchical procedure set out above. We adopt a uniform, improper hyperprior on both, and use a diagonal $\mathbf{C_d^{-1}}$ constructed using the uncertainties associated with Hoggard's point-wise measurements.

Applying eq. (17), we first map out the distribution $\mathbb{P}[\alpha, \beta \mid \mathbf{d_0}]$ in the vicinity of its maximum, as shown in Fig. 7. Again, we see a well-defined, tightly constrained maximum, which lies at the point $\alpha_0 \approx 3.3$, $\beta_0 \approx 1.2$. Approximating the distribution by a δ-distribution at this point, we obtain the best-fitting model coefficients, and compute the power in this at the lth degree, $P_l = \sum_{m=-l}^{l} c_{lm}^2$. The resulting spectrum (blue), and that obtained by Hoggard (black), are shown in Fig. 8, which follows the presentation in Fig. (5b) of Hoggard *et al.* (2016). To provide some estimate of the uncertainty associated with this result, we also show the range spanned by spectra for 100 000 models generated at random from our posterior distribution, $\mathbb{P}[\mathbf{m} \mid \mathbf{d_0}, \alpha_0, \beta_0]$; the spectral distribution is non-Gaussian, so 'standard' error analysis

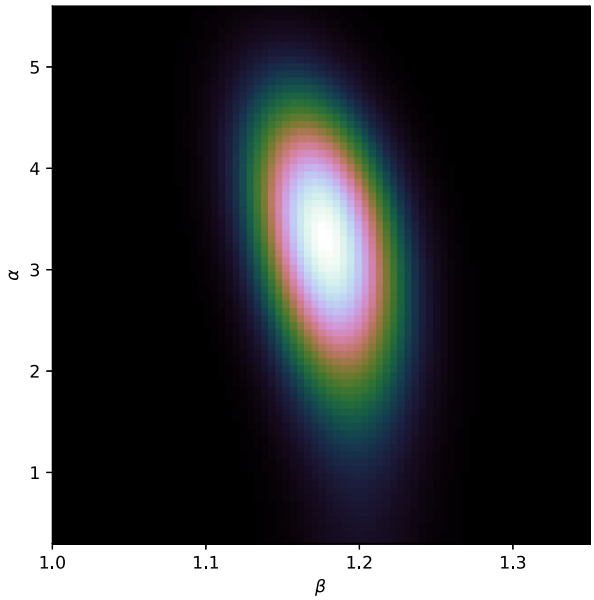

Figure 7. The distribution $\mathbb{P}[\alpha, \beta \mid \mathbf{d_0}]$ for spherical harmonic expansion of dynamic topography estimates (as per Hoggard *et al.* 2016).

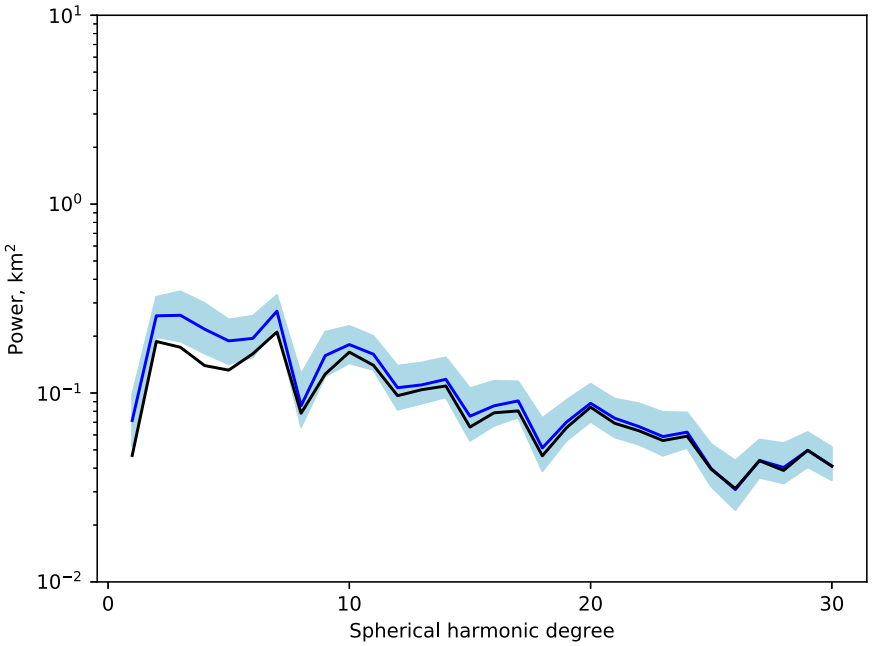

Figure 8. Power spectra for dynamic topography; compare Fig. 5(b) of Hoggard *et al.* (2016). Black: results reported by Hoggard (using his preferred choice of regularization). Dark blue: results obtained using optimal regularization determined using the techniques of this paper. Light blue: range of spectra for 100 000 samples from posterior distribution. Although we find that Hoggard's regularization choices were sub-optimal (according to the definitions of this paper), this does not significantly alter his result.

may mislead. Using our optimal choice of regularization, we obtain results that are very similar to those reported in Hoggard *et al.* (2016). Thus, it does not appear that his reported discrepancies against simulation results (which predict power of 1–3 km^2 at the longest wavelengths) can be resolved simply through a 'better' choice of regularization. Nevetheless, this example demonstrates that our approach can be effective in realistic-scale problems; some discussion of computational costs can be found in Section 5.5.

5 DISCUSSION

The key points from Section 2 may be summarized as follows:

(i) Any prior we adopt in model space implicitly defines the range within which we expect observations to lie. Both model and data space should be taken into account when assessing whether a given prior is 'reasonable'.

(ii) For a general parametric family of model covariance matrices, $\mathbf{C_m}(\boldsymbol{\xi})$, the probability distribution $\mathbb{P}[\boldsymbol{\xi} \mid \mathbf{d_0}]$ may be evaluated and used to assess the inherent plausibility of different values of $\boldsymbol{\xi}$.

(iii) The distribution $\mathbb{P}[\boldsymbol{\xi} \mid \mathbf{d_0}]$ is seen to typically have a well-defined maximum-likelihood point, which can, in general, be found by setting up a non-linear optimization problem.

(iv) In principle, the full solution to the inverse problem can be found by integrating over the space of all regularization parameters. However, it may be reasonable to approximate $\mathbb{P}[\boldsymbol{\xi} \mid \mathbf{d_0}]$ by a simpler distribution. If a δ-distribution is used, the procedure can be viewed as equivalent to identification of an 'optimal' set of regularization parameters to use in conjunction with the standard analysis (eq. 9).

As demonstrated in the previous section, this hierarchical approach can yield results that are in broad agreement with those obtained using alternative strategies for selection of regularization parameters—but with less dependence upon subjective choices made by the user.

5.1 The significance of the maximum-likelihood regularisers

The use of $\mathbb{P}[\boldsymbol{\xi} \mid \mathbf{d_0}]$ to determine the correct regularization term is a natural outcome of the hierarchical Bayesian framework, but may appear counter-intuitive. To gain insight into the effect of the hierarchical procedure, it is helpful to consider the properties of the regulariser at the maximum-likelihood point (i.e. the peak of $\mathbb{P}[\boldsymbol{\xi} \mid \mathbf{d_0}]$).

5.1.1 The general case

First, we consider the most general covariance matrix, $\mathbf{C_m}$, with all elements of the matrix being independently determined (which may of course be treated within the general parametric framework of Section 3.1). We note that if a vector $\mathbf{x}$ is normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$, it can be shown that $\mathbb{E}\left[\mathbf{x}\mathbf{x}^\mathbf{T}\right] = \mathbf{C} + \boldsymbol{\mu}\boldsymbol{\mu}^\mathbf{T}$. Thus, using the results from eq. (9), eq. (19) can be expressed in the form

$$\frac{\partial \log \mathbb{P}\left[\mathbf{d}' \mid \mathbf{C_m^{-1}}\right]}{\partial \mathbf{C_m^{-1}}} = \frac{1}{2}\left\{\mathbf{C_m} - \tilde{\mathbb{E}}\left[(\mathbf{m}-\mathbf{m_p})(\mathbf{m}-\mathbf{m_p})^\mathbf{T}\right]\right\}, \tag{37}$$

where a tilde denotes the expectation with respect to the posterior distribution (which, of course, itself depends upon $\mathbf{C_m}$). Recalling the general definition of covariance, $\mathbb{C}[\mathbf{x}] = \mathbb{E}\left[\left(\mathbf{x}-\mathbb{E}[\mathbf{x}]\right)\left(\mathbf{x}-\mathbb{E}[\mathbf{x}]\right)^\mathbf{T}\right]$, we can regard the second term in eq. (37) as an estimate of the posterior covariance, $\boldsymbol{\Psi}$, obtained under the assumption that the posterior is indeed centred upon $\mathbf{m_p}$. Then, since eq. (37) must be equal to zero at the optimum, the preferred covariance matrix is one that satisfies

$$\mathbf{C_m} = \tilde{\mathbb{E}}\left[(\mathbf{m}-\mathbf{m_p})(\mathbf{m}-\mathbf{m_p})^\mathbf{T}\right] - 2\frac{\partial \log \mathbb{P}\left[\mathbf{C_m^{-1}}\right]}{\partial \mathbf{C_m^{-1}}}. \tag{38}$$

If we have no strong preferences regarding the structure of the matrix $\mathbf{C_m^{-1}}$, it may be appropriate to adopt a uniform hyperprior. In this case, the second term in eq. (38) disappears, and thus the 'optimal' model covariance matrix is the one that only changes between prior and posterior distributions if the centre of the distribution also changes. In effect, then, the hierarchical procedure amounts to estimation of the model covariance structure supported by the data, under the assumption that $\mathbf{m_p}$ correctly identifies the most-probable model. This ensures that all assumptions are self-consistent throughout the analysis.

Another route to this understanding comes by introducing the concept of a 'resolution operator' (Backus & Gilbert 1968). If we imagine computing synthetic data, $\mathbf{s} = \mathbf{G}\mathbf{m}_{\text{true}}$, and then treating it as the data for inversion according to eq. (9), we obtain $\mathbf{m}_{\text{est}} = \mathbf{R}\mathbf{m}_{\text{true}}$ with

$$\mathbf{R} = \left(\mathbf{G}^\mathbf{T}\mathbf{C_d^{-1}}\mathbf{G} + \mathbf{C_m^{-1}}\right)^{-1}\mathbf{G}^\mathbf{T}\mathbf{C_d^{-1}}\mathbf{G}. \tag{39}$$

This 'resolution matrix'—which depends entirely upon the imaging setup, and not on the data—can therefore be regarded as a 'filter' through which recovered models perceive reality. This allows us to write eq. (19) in yet another form,

$$\frac{\partial \log \mathbb{P}\left[\mathbf{d}' \mid \mathbf{C_m^{-1}}\right]}{\partial \mathbf{C_m^{-1}}} = \frac{1}{2}\left\{\mathbf{R}\mathbf{C_m} - \left(\mathbf{m}_{\text{est}}-\mathbf{m_p}\right)\left(\mathbf{m}_{\text{est}}-\mathbf{m_p}\right)^\mathbf{T}\right\}. \tag{40}$$

Here, both $\mathbf{R}$ and $\mathbf{m}_{\text{est}}$ should be understood to be implicitly dependent upon $\mathbf{C_m^{-1}}$, and we note that $\mathbf{R}^\mathbf{T} = \mathbf{C_m^{-1}}\mathbf{R}\mathbf{C_m}$, and so $\partial \log \mathbb{P}\left[\mathbf{d}' \mid \mathbf{C_m^{-1}}\right]/\partial \mathbf{C_m^{-1}}$ is symmetric, as required. Setting the derivative equal to zero, we see that the search for an optimal covariance matrix is equivalent to tuning the imaging process until the recovered model (relative to the mean of the prior) satisfies the eigenvector equation

$$\left(\mathbf{R} + 2\frac{\partial \log \mathbb{P}\left[\mathbf{C_m^{-1}}\right]}{\partial \mathbf{C_m^{-1}}}\mathbf{C_m^{-1}}\right)\left(\mathbf{m}_{\text{est}}-\mathbf{m_p}\right) = \kappa\left(\mathbf{m}_{\text{est}}-\mathbf{m_p}\right), \tag{41a}$$

with eigenvalue given by

$$\kappa = (\mathbf{m}_{\text{est}}-\mathbf{m_p})^\mathbf{T}\mathbf{C_m^{-1}}(\mathbf{m}_{\text{est}}-\mathbf{m_p}). \tag{41b}$$

Again, if we lack strong information about $\mathbf{C_m^{-1}}$, and hence adopt a flat hyperprior, the term involving $\mathbb{P}\left[\mathbf{C_m^{-1}}\right]$ can be neglected. Choosing $\mathbf{C_m^{-1}}$ to maximise eq. (18) therefore ensures that the recovered model is an eigenvector of the resolution matrix. This represents a 'stable'

inverse operator: if a synthetic data set for $\mathbf{m}_{\text{est}}$ is inverted, the recovered model is unchanged (except for an amplitude scaling, which can be readily corrected). Again, this may be viewed as a requirement for self-consistency. Unfortunately, there are no corresponding implications for the accuracy of $\mathbf{m}_{\text{est}}$ relative to the 'true' (data-generating) model, since $\mathbf{R}$ is not generally symmetric and hence its eigenvectors need not span the model space. It is perhaps worth stressing that the property stated in eq. (41) only applies in the case where no restrictions have been placed upon the form of $\mathbf{C}_{\mathbf{m}}^{-1}$.

While these self-consistency properties are attractive, it will readily be appreciated that estimating model covariance structure from a single data vector is a potential source of difficulty. In particular, if any of the assumptions underpinning the Bayesian analysis are not fully met (perhaps due to an imperfect forward theory, or non-Gaussian noise), or where the data set is such that it provides an uneven sampling of the model parameters, a biased estimate of the covariance structure may result. A wealth of statistical research has focussed on the problem of developing unbiased estimators for particular quantities, and it may be possible to use these strategies within our regularization framework. However, exploring this question is beyond the scope of the current paper. An effective—though undoubtedly less attractive—alternative is to use the hyperprior, $\mathbb{P}\left[\mathbf{C}_{\mathbf{m}}^{-1}\right]$, to ensure choices remain within some 'sensible' range. In such circumstances, eqs (38) and (41) reveal that the data-derived estimated covariance is tempered by some 'preferred' structure encoded within the hyperprior.

5.1.2 Tikhonov-style regularization

As already emphasized, the above results do not hold if the optimization process is constrained by a requirement that $\mathbf{C}_{\mathbf{m}}$ conform to a specified form. In any such case, further analysis is required to take account of the additional restrictions imposed upon $\mathbf{C}_{\mathbf{m}}$. We therefore now consider the Tikhonov-style regularization discussed in Section 3.2. Again, we begin by noting that for $\mathbf{x}$ again Gaussian-distributed with mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$, and for a general matrix $\mathbf{A}$, it can be shown that $\mathbb{E}\left[\mathbf{x}^{\mathbf{T}}\mathbf{A}\mathbf{x}\right] = \operatorname{Tr}\mathbf{A}\mathbf{C} + \boldsymbol{\mu}^{\mathbf{T}}\mathbf{A}\boldsymbol{\mu}$. Thus, eq. (23) may be rewritten as

$$\frac{\partial \log \mathbb{P}\left[\alpha, \beta \mid \mathbf{d}'\right]}{\partial \alpha} = \alpha \left\{ \mathbb{E}\left[\left(\mathbf{m}-\mathbf{m}_{\mathbf{p}}\right)^{\mathbf{T}} \left(\mathbf{m}-\mathbf{m}_{\mathbf{p}}\right) \right] - \tilde{\mathbb{E}}\left[\left(\mathbf{m}-\mathbf{m}_{\mathbf{p}}\right)^{\mathbf{T}} \left(\mathbf{m}-\mathbf{m}_{\mathbf{p}}\right) \right] \right\} + \frac{\partial \log \mathbb{P}\left[\alpha, \beta\right]}{\partial \alpha} \tag{42a}$$

and

$$\frac{\partial \log \mathbb{P}\left[\alpha, \beta \mid \mathbf{d}'\right]}{\partial \beta} = \beta \left\{ \mathbb{E}\left[\left(\mathbf{m}-\mathbf{m}_{\mathbf{p}}\right)^{\mathbf{T}} \mathbf{H} \left(\mathbf{m}-\mathbf{m}_{\mathbf{p}}\right) \right] - \tilde{\mathbb{E}}\left[\left(\mathbf{m}-\mathbf{m}_{\mathbf{p}}\right)^{\mathbf{T}} \mathbf{H} \left(\mathbf{m}-\mathbf{m}_{\mathbf{p}}\right) \right] \right\} + \frac{\partial \log \mathbb{P}\left[\alpha, \beta\right]}{\partial \beta} . \tag{42b}$$

If we wish to determine α and β purely from the data, without expressing any preferences regarding their values, we would adopt a uniform hyperprior $\mathbb{P}\left[\alpha, \beta\right]$, and hence the final term in these expressions would disappear. We remark that in this case, trivial extrema of $\mathbb{P}\left[\alpha, \beta \mid \mathbf{d}'\right]$ exist at $\alpha = 0$ and $\beta = 0$; in practice, these are generally minima, although this may be verified by checking the second derivative via the expressions given in Appendix A1. It is not possible for both α and β to be zero, as the result would no longer be a valid covariance matrix.

Apart from this trivial solution, we see that optimization of α and β drives the system towards a state where the expected model norm, and expected model 'smoothness', only change from prior to posterior if the mean of the distribution itself changes. The optimization procedure essentially amounts to estimating these quantities from the data; then, this information is used to 'constrain' the inversion (in a loosely defined sense). Again, this is desirable from a 'self-consistency' perspective: since the prior implicitly encodes a preference for models of a certain size and smoothness, any change in the expected values of these quantities from prior to posterior may be regarded as an indication that our initial estimates were inadequate. In the event that we have a preference for model norms or smoothnesses of a certain size, a non-uniform hyperprior on α and β may be used to impart a bias into the estimation procedure. Depending on circumstance, a variety of choices of hyperprior may be appropriate: common options might include independent gamma distributions on α and β, or a joint normal distribution.

5.2 Priors and hyperpriors

In Section 2.4, we approached the hierarchical procedure via a prior over both model parameters and hyperparameters, $\mathbb{P}\left[\mathbf{m}, \boldsymbol{\xi}\right]$. Our strategy for solving the inverse problem is then, in effect, to first find $\mathbb{P}\left[\mathbf{m}, \boldsymbol{\xi} \mid \mathbf{d}_{\mathbf{0}}\right]$, and then marginalize out any dependence on $\boldsymbol{\xi}$. Formally, however, it is possible to interchange the order of these operations: we can marginalise over $\boldsymbol{\xi}$, to leave a prior defined over $\mathbf{m}$ only, and then condition this upon the observed data. This is impractical as a solution strategy—we would lose the ability to exploit the analytic results underpinning eq. (9)—but is instructive for understanding the role of the hierarchical procedure.

For the example depicted in Figs 3 and 4, we effectively adopted

$$\mathbb{P}\left[\mathbf{m}, \alpha\right] = \begin{cases} \frac{\alpha^M}{(2\pi)^{M/2}} \exp\left(-\frac{\alpha^2}{2}\mathbf{m}^{\mathbf{T}}\mathbf{m}\right) & 0 < \alpha \leq 10 \\ 0 & \text{otherwise.} \end{cases}$$

Integrating over α, we can obtain $\mathbb{P}\left[\mathbf{m}\right]$; noting that the distribution is symmetric in all M model parameters, a 1-D marginal is shown in Fig. 9. It can be seen that this has significantly heavier tails than a Gaussian distribution. This implies that, compared to a Gaussian prior of equivalent width, the hierarchical procedure expresses greater willingness to move substantially away from the *a priori* most probable model, $\mathbf{m}_{\mathbf{p}}$.

The precise form of $\mathbb{P}\left[\mathbf{m}\right]$ within the hierarchical procedure is governed by the choice of hyperprior $\mathbb{P}\left[\boldsymbol{\xi}\right]$. In principle, this should be chosen to reflect existing knowledge about the range of plausible solutions to the inverse problem; in practice, the choice will often be made

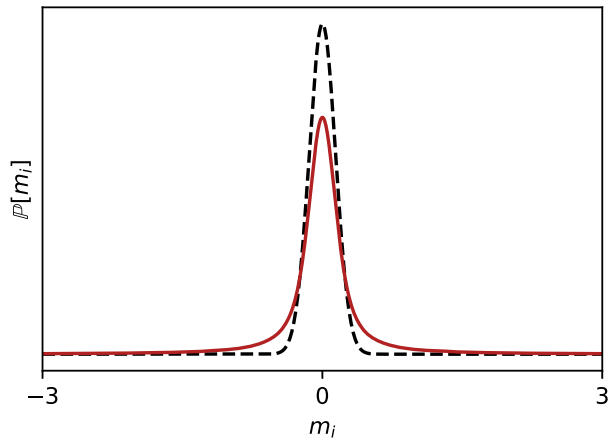

Figure 9. The prior distribution (red) effectively used for the construction of Figs 3 and 4. For comparison, a Gaussian distribution of similar width is shown as a dashed black line; it can be seen that the prior has significantly heavier tails than the Gaussian.

on an *ad hoc* basis. However, experience suggests that for many problems, the solution (i.e. the posterior distribution) is relatively insensitive to the details of the hyperprior. Thus, the precise choice made is not critical to obtaining meaningful results. Nevertheless, the range of 'reasonable' choices may be problem-dependent, and we do not provide specific advice here.

For certain choices of hyperprior, such as a gamma distribution, it may be possible to solve the integral of eq. (10) analytically. In this case, $\mathbb{P}[\boldsymbol{\xi}]$ is said to be 'conjugate' to the Gaussian distribution, $\mathbb{P}[\mathbf{m} \,|\, \boldsymbol{\xi}]$. The analysis of conjugate distributions was a central pillar of Bayesian statistics before computational techniques such as Monte Carlo algorithms became widely accessible, and a number of powerful results can be found (e.g Fink 1997). In some cases, it may be possible to perform all analysis analytically, with the posterior distribution $\mathbb{P}[\mathbf{m} \,|\, \mathbf{d_0}]$ taking the form of a known family of distributions (generally non-Gaussian). However, the need to map between model- and data-spaces, and to accommodate data noise, introduces difficulties that would appear to prevent straightforward application of standard results.

5.3 Non-linear inversion

This paper has focussed entirely on linear inverse problems, where the data are assumed to be a linear function of the model parameters. For many problems of geophysical interest this is not the case. Nevertheless, if the non-linearity is relatively weak, it is possible to apply the least-squares algorithm iteratively and converge upon a solution to the inverse problem. Again, the difficulties of choosing a prior (or, equivalently, defining an appropriate regularization scheme) arise: can the analysis of this paper be used?

It is difficult to give a universal answer to this question. Our approach relies on using the linear operator, $\mathbf{G}$, to map the prior distribution into the model space. If non-linearity is sufficiently weak, and/or the prior sufficiently narrow, this may continue to perform well in non-linear problems. However, this will not hold for all priors and all forward models, and some degree of case-by-case investigation may be required. We remark that some assistance may be gained by following Tarantola & Valette (1982) and introducing a Gaussian 'theory error', used to artificially inflate the data noise. We also note that in the non-linear case, the matrix $\mathbf{G}$ appearing in various expressions in this paper should be the linear operator corresponding to the point $\mathbf{m_p}$: the prior should remain fixed over multiple iterations, and not be updated as the inversion proceeds.

5.4 Non-Bayesian inversion

The analysis of this paper has been framed in terms of Bayes' theorem, and builds on the work of Tarantola & Valette (1982). Can our results be used for the regularization of inverse problems framed 'deterministically', that is without introducing the machinery of probability calculus? At least in a purely mechanical sense, this appears to be the case: if we replace $\mathbf{C_m^{-1}}$ by a damping matrix $\mathbf{D}$, and introduce a weighting matrix to substitute for $\mathbf{C_d^{-1}}$, the algorithms to find the maximum-likelihood regulariser can be used to select a choice of $\mathbf{D}$ that is 'optimal' (in a sense that could perhaps be formalised through the properties highlighted in Section 5.1). As we have seen, these maximum-likelihood regularization parameters are often effective for obtaining good results. Thus, the results of this paper could be applied to problems established following the analysis set out in Section 2.1.

5.5 Software implementation and practical considerations

Python routines implementing the main ideas of this paper are provided as Supplementary Material.[1] In particular, we provide solvers for optimally regularized least squares (i.e. using the approximation $\mathbb{P}[\boldsymbol{\xi} \mid \mathbf{d_0}] \approx \delta(\boldsymbol{\xi} - \boldsymbol{\xi_0})$) for both general parametric covariance matrices, $\mathbf{C_m}(\boldsymbol{\xi})$, and for Tikhonov-style regularization. Our implementation uses the `L_BFGS_B` optimization algorithm (Liu & Nocedal 1989) to perform maximization of $\mathbb{P}[\mathbf{d_0} \mid \boldsymbol{\xi}]$ in the general case, with Brent's method being adopted for root-finding using eqs (26) and (28). Other approaches may well be possible, and we do not claim that our implementation is necessarily optimal in any sense. We envisage that end-users will wish to adapt and enhance our routines to suit the needs of their applications. To provide some illustration of the software 'in action', code to generate the examples of Section 4 is also provided.

In the case of Tikhonov regularization where both α and β parameters are allowed to vary, two different strategies may be adopted. First—as in the accompanying routines—one may simply apply the theory for general covariance matrices; alternatively, one can adopt an iterative scheme alternating between optimization of α for fixed β using eq. (26), and optimization of β for fixed α via eq. (28). Our experience suggests that the latter strategy is substantially more efficient than the former, but less stable. This is likely to be problem-dependent, and users may wish to explore performance for their particular problems of interest.

We emphasise that the procedure of determining optimal regularization parameters via eqs (26) and (28) is computationally efficient, with only modest additional overhead beyond that required to solve the inverse problem for a pre-determined choice of regularization. The principal computational costs are those associated with performing eigendecomposition on the various matrices; once this is done, the root-finding procedure is relatively cheap. For general covariance matrices, costs are higher, due to the need to re-invert $(\mathbf{G^T C_d^{-1} G} + \mathbf{C_m^{-1}}(\boldsymbol{\xi}))$ at every iteration of the optimization procedure. For problems of more than modest dimension, this may be prohibitive; however, knowledge of the specific form of $\mathbf{C_m}(\boldsymbol{\xi})$ may enable more efficient implementations to be developed.

To make this more concrete, we return to the dynamic topography example of Section 4.3. This inverse problem has 960 model parameters; if the regularization is fixed *a priori*, an inversion costs around 0.2 CPU-seconds on a standard desktop PC (an Apple iMac with a 3.2 GHz processor). If α is not known, and is to be determined using eq. (26), the total cost increases to around 0.5 CPU-seconds. If both α and β must be found, the total execution time is 13 CPU-seconds, although multithreading can reduce the time requirement from a user's perspective. Although a substantial increase on the cost of an inversion for fixed α and β, this still compares favourably to the effort required to produce an L-curve, or similar, and select appropriate regularization choices. Some reduction in costs could be achieved—at the expense of generality—by adapting the algorithm to exploit the properties of the particular choice of $\mathbf{H}$ adopted. This may be worthwhile for large-scale problems, such as tomographic inversions, where the dimension of the model-space may be substantially larger than in this example, and where the cost of matrix inversion is therefore much higher. However, the details of any potential optimizations must inevitably be determined on a case-by-case basis.

6 CONCLUDING REMARKS

This paper has developed a hierarchical Bayesian approach to specification of the prior model covariance matrix in linear inverse problems set within the probabilistic framework of Tarantola & Valette (1982). Equivalently, this can be seen as an extension of the work of Tarantola & Valette (1982) to a certain class of non-Gaussian, heavy-tailed prior distribution. Our results may also be interpreted as providing a method for determination of the 'optimal' regularization parameters to use in deterministic inversion. The regularization choices indicated by our approach are broadly similar to those suggested by established methods, such as picking the 'elbow' on an L-curve—but we are able to avoid the need for any subjective input from the user. This increases the rapidity with which results may be obtained, promotes reproducibility, and—we hope—may help minimise controversy surrounding the choices made.

In this paper, we have assumed that a single *a priori* most-likely model, $\mathbf{m_p}$ can be identified, so that only the covariance matrix of the prior needs to be treated as an unknown. In many geophysical problems, this will be a reasonable restriction. However, in principle it is possible to extend the hierarchical approach to also include uncertainty around the correct value $\mathbf{m_p}$, and this may be an avenue amenable to further study. Equivalently, this may be seen as extending the analysis of Tarantola & Valette (1982) to encompass a wider variety of non-Gaussian priors.

Access to an automated method for selecting regularization parameters makes the use of more complex regularization schemes feasible. In particular, the concept of Automatic Relevance Determination introduced by Mackay (1992a) appears to have powerful application in geophysical inverse problems, and deserves further investigation. Recent work has demonstrated that sparsity can be a powerful property to exploit in inference problems, but numerical optimization under L_0 or L_1 constraints can be challenging. Does the Automatic Relevance Determination procedure offer a route to recovering sparse models within an L_2 framework? Our results in Section 4 would appear to suggest that this is indeed the case, although substantial further investigation is required. Nevertheless, given the success of compressive sensing and related techniques, we suggest that this may be a fruitful avenue to pursue.

[1] Also available via a Git repository at http://github.com/valentineap/optimal-regularisation/

ACKNOWLEDGEMENTS

We thank Mark Hoggard for providing the data set used for Section 4.3, and for useful discussions. Conversations with David Al-Attar, Rhodri Davies and Jeannot Trampert have also been valuable in this work. APV acknowledges support from the Australian Research Council through a Discovery Early Career Research Award (grant number DE180100040), from Geoscience Australia (under the auspices of the project "Data Science in Solid Earth Geophysics"), and from the Research School of Earth Sciences at ANU. Finally, we are grateful to Clifford Thurber, an anonymous reviewer, and the editor, Frederik Simons, for their constructive comments upon the original version of this manuscript.

REFERENCES

Abramowitz, M. & Stegun, I., 1964. *Handbook of Mathematical Functions, no. 55 in Applied Mathematics Series,* National Bureau of Standards.

Allen, D., 1974. The relationship between variable selection and data augmentation and a method for prediction, *Technometrics,* **16,** 125–127.

Aster, R., Borchers, B. & Thurber, C., 2013. *Parameter Estimation and Inverse Problems,* Academic Press.

Backus, G. & Gilbert, F., 1968. The resolving power of gross Earth data, *Geophys. J. R. astr. Soc.,* **16,** 169–205.

Bayes, T., 1763. An essay towards solving a problem in the doctrine of chances, *Philos. Trans.,* **53,** 370–418.

Bernardi, F., Braunmiller, J., Kradolfer, U. & Giardini, D., 2004. Automatic regional moment tensor inversion in the European-Mediterranean region, *Geophys. J. Int.,* **157,** 703–716.

Boschi, L. & Dziewoński, A., 1999. High- and low-resolution images of the Earth's mantle: implications of different approaches to tomographic modeling, *J. geophys. Res.,* **104,** 25567–25594.

Candès, E. & Wakin, B., 2008. An introduction to compressive sampling, *IEEE Signal Proc. Mag.,* **25,** 21–30.

Constable, S., Orange, A. & Key, K., 2015. And the geophysicist replied: "which model do you want?", *Geophysics,* **80,** E197–E212.

Deuss, A., Ritsema, J. & van Heijst, H., 2013. A new catalogue of normal-mode splitting function measurements up to 10 mHz, *Geophys. J. Int.,* **193,** 920–937.

Donoho, D., 2006a. For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution, *Commun. Pure appl. Math.,* **59,** 797–829.

Donoho, D., 2006b. For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution, *Commun. Pure appl. Math.,* **59,** 907–934.

Dziewonski, A., Chou, T.-A. & Woodhouse, J., 1981. Determination of earthquake source parameters from waveform data for studies of global and regional seismicity, *J. geophys. Res.,* **86,** 2825–2852.

Fink, D., 1997. *A Compendium of Conjugate Priors,* Tech. rep, Montana State University.

Golub, G., Heath, M. & Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics,* **21,** 215–223.

Hansen, P., 1992. Analysis of discrete ill-posed problems by means of the L-curve, *SIAM Rev.,* **34,** 561–580.

Hansen, P. & O'Leary, D., 1993. The use of the L-curve in the regularization of discrete ill-posed problems, *SIAM J. Scient. Comput.,* **14,** 1487–1503.

Hoggard, M., White, N. & Al-Attar, D., 2016. Global dynamic topography observations reveal limited influence of large-scale mantle flow, *Nature Geosci.,* **9,** 456–463.

Lekić, V. & Romanowicz, B., 2011. Inferring mantle structure by full waveform tomography using the spectral element method, *Geophys. J. Int.,* **185,** 799–831.

Liu, D. & Nocedal, J., 1989. On the limited memory BFGS method for large-scale optimization, *Math. Program.,* **45,** 503–528.

Mackay, D., 1992a. Bayesian interpolation, *Neural Comput.,* **4,** 415–447.

Mackay, D., 1992b. A practical Bayesian framework for backpropagation networks, *Neural Comput.,* **4,** 448–472.

Menke, W., 1989. *Geophysical Data Analysis: Discrete Inverse Theory,* Academic Press, New York.

Morozov, V., 1968. The error principle in the solution of operational equations by the regularization method, *USSR Comput. Math. Math. Phys.,* **8,** 63–87.

Pavlis, N., Holmes, S., Kenyon, S. & Factor, J., 2012. The development and evaluation of the Earth Gravitational Model 2008 (EGM2008), *J. geophys. Res.,* **117,** B04406.

Petersen, K. & Pedersen, M., 2015. *The matrix cookbook,* Tech. rep., Technical University of Denmark.

Pratt, R. & Worthington, M., 1990. Inverse theory applied to multi-source cross-hole tomography. Part 1: acoustic wave-equation method, *Geophys. Prospect.,* **38,** 287–310.

Rawlinson, N., Sambridge, M. & Saygin, E., 2008. A dynamic objective function technique for generating multiple solution models in seismic tomography, *Geophys. J. Int.,* **174,** 295–308.

Ritsema, J., Deuss, A., van Heijst, H. & Woodhouse, J., 2011. S40RTS: a degree-40 shear-velocity model for the mantle from new Rayleigh wave dispersion, teleseismic traveltime and normal-mode splitting function measurements, *Geophys. J. Int.,* **184,** 1223–1236.

Schaeffer, A. & Lebedev, S., 2015.Global heterogeneity of the lithosphere and underlying mantle: A seismological appraisal based on multimode surface-wave dispersion analysis, shear-velocity tomography, and tectonic regionalization, in *The Earth's heterogeneous mantle,* pp. 3–46, in Khan, A. & Deschamps, F., eds., Springer.

Spakman, W., van der Lee, S. & van der Hilst, R., 1993. Travel-time tomography of the European-Mediterranean mantle down to 1400 km, *Phys. Earth planet. Inter.,* **79,** 3–74.

Tarantola, A. & Valette, B., 1982. Generalized nonlinear inverse problems solved using the least squares criterion, *Rev. Geophys. Space Phys.,* **20,** 219–232.

Valentine, A. & Trampert, J., 2016. The impact of approximations and arbitrary choices on geophysical images, *Geophys. J. Int.,* **204,** 59–73.

Woodhouse, J. & Dziewonski, A., 1984. Mapping the upper mantle: three-dimensional modelling of Earth structure by inversion of seismic waveforms., *J. geophys. Res.,* **89,** 5953–5986.

Yagi, Y. & Fukahata, Y., 2011. Introduction of uncertainty of Green's function into waveform inversion for seismic source processes, *Geophys. J. Int.,* **186,** 711–720.

SUPPORTING INFORMATION

Supplementary data are available at *GJI* online.

optimal-regularization-1.1.tar

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX: MISCELLANEOUS FORMULAE

This appendix tabulates certain expressions that may be useful for computational implementation of the foregoing theory, but which were not otherwise found useful for its explanation. Again, we commend the work of Petersen & Pedersen (2015) to readers who wish to derive or manipulate these formulae. It should be assumed that all quantities are defined as in the main text.

A1 Second derivatives for Tikhonov-style regularization

Differentiating the expressions in eq. (23) a second time yields

$$\frac{\partial^2 \log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right]}{\partial \alpha^2} = \operatorname{Tr}\left\{\left(\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-1} - \left(\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}+\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-1}\right\} - 2\alpha^2 \operatorname{Tr}\left\{\left(\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-2} - \left(\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}+\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-2}\right\}$$
$$-\alpha\mathbf{d}'^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}\left(\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}+\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-2}\left[\mathbf{I}-4\alpha\left(\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}+\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-1}\right]\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{d}' \tag{A1a}$$

$$\frac{\partial^2 \log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right]}{\partial \beta^2} = \operatorname{Tr}\left\{\mathbf{H}\left[\left(\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-1} - \left(\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}+\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-1}\right]\right\}$$
$$-\beta\mathbf{d}'^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}\left(\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}+\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-1}\mathbf{H}\left(\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}+\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-1}[\mathbf{I}$$
$$-4\beta\mathbf{H}\left(\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}+\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-1}\Big]\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{d}' - 2\beta^2 \operatorname{Tr}\left\{\mathbf{H}\left[\left(\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-1}\mathbf{H}\left(\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-1}\right.\right.$$
$$\left.\left.-\left(\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}+\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-1}\mathbf{H}\left(\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}+\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-1}\right]\right\} \tag{A1b}$$

$$\frac{\partial^2 \log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right]}{\partial \alpha \partial \beta} = 2\alpha\beta \operatorname{Tr}\left\{\mathbf{H}\left[\left(\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-2} - \left(\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}+\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-2}\right]\right\}$$
$$+4\alpha\beta\mathbf{d}'^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}\left(\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}+\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-1}\mathbf{H}\left(\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G}+\alpha^2\mathbf{I}+\beta^2\mathbf{H}\right)^{-2}\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{d}' \tag{A1c}$$

Together, these expressions define the Hessian for $\log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right]$, and (assuming a uniform prior on α and β) they may be used to determine the covariance matrix for a Gaussian approximation to $\mathbb{P}\left[\alpha, \beta \mid \mathbf{d}'\right]$.

A2 Tikhonov-style regularization, fixed β

In Section 3.2.1, we assume $\mathbf{C}_{\mathbf{m}}^{-1} = \alpha^2\mathbf{I} + \beta^2\mathbf{H}$ for fixed β, and obtain an efficient algorithm for estimating the corresponding optimum value of α_0 in terms of the eigendecompositions $\mathbf{H} = \mathbf{T}\boldsymbol{\Omega}\mathbf{T}^{\mathbf{T}}$ and $\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G} + \beta^2\mathbf{H} = \mathbf{S}\boldsymbol{\Lambda}\mathbf{S}^{\mathbf{T}}$. Making these substitutions, eq. (18) becomes

$$\log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right] = \frac{1}{2}\left\{\left(\log\left|\mathbf{C}_{\mathbf{d}}^{-1}\right| - N\log 2\pi - \mathbf{d}'^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{d}'\right) + \sum_{i=1}^{M}\left[\log\left(\alpha^2+\beta_2\omega_i\right) - \log\left(\lambda_i+\alpha^2\right) + \frac{\left[\mathbf{S}^{\mathbf{T}}\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{d}'\right]_i^2}{\lambda_i+\alpha^2}\right]\right\} \tag{A2}$$

Again, we note that if $\beta = 0$, the first term in the sum must nevertheless be included M times, to give a total contribution $2M\log\alpha$. The second derivative with respect to α, from eq. (A1), is expressed

$$\frac{\partial^2 \log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right]}{\partial \alpha^2} = \sum_{i=1}^{M}\left[\frac{\beta^2\omega_i - \alpha^2}{\left(\alpha^2+\beta^2\omega_i\right)^2} + \frac{\alpha^2-\lambda_i}{\left(\lambda_i+\alpha^2\right)^2} - \frac{\alpha^3 - 4\alpha^2 + \alpha\lambda_i}{\left(\lambda_i+\alpha^2\right)^3}\left[\mathbf{S}^{\mathbf{T}}\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{d}'\right]_i^2\right]. \tag{A3}$$

Since we are regarding β as a constant, this is the only non-zero second derivative. Recalling that eq. (18) had a trivial extremum at $\alpha = 0$, we remark that $\partial^2 \log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right]/\partial\alpha^2\big|_{\alpha=0}$ is negative if

$$\sum_{i=1}^{M}\left[\frac{1}{\beta^2\omega_i} - \frac{1}{\lambda_i}\right] < 0 \tag{A4}$$

and hence $\alpha = 0$ corresponds to a probability maximum only if this condition is met.

A3 Tikhonov-style regularization, fixed α

Similarly, in Section 3.2.2 we consider the case where α is fixed, and β is allowed to vary. Using $\mathbf{H}^{-1}\left(\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{G} + \alpha^2\mathbf{I}\right) = \mathbf{U}\boldsymbol{\Gamma}\mathbf{U}^{-1}$, eq. (18) can be written

$$\log \mathbb{P}\left[\mathbf{d}' \mid \alpha, \beta\right] = \frac{1}{2}\left\{\left(\log\left|\mathbf{C}_{\mathbf{d}}^{-1}\right| - N\log 2\pi - \mathbf{d}'^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{d}'\right)\right.$$
$$\left. + \sum_{i=1}^{M}\left[\log\left(\alpha^2+\beta^2\omega_i\right) - \log\omega_i - \log\left(\gamma_i+\beta^2\right) - \frac{\left[\mathbf{U}^{\mathbf{T}}\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{d}'\right]_i\left[\mathbf{U}^{-1}\mathbf{H}^{-1}\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-1}\mathbf{d}'\right]_i}{\left(\gamma_i+\beta_0^2\right)}\right]\right\} \tag{A5}$$

and the only non-zero second derivative is

$$\frac{\partial^2 \log \mathbb{P}\left[\mathbf{d}' \,|\, \alpha, \beta\right]}{\partial \beta^2} = \sum_{i=1}^{M} \left[\frac{\omega_i \left(\alpha^2 - \beta^2 \omega_i\right)}{\left(\alpha^2 + \beta^2 \omega_i\right)^2} + \frac{\beta^2 - \gamma_i}{\left(\gamma_i + \beta^2\right)^2} - \frac{\beta^3 - 4\beta^2 + \beta\gamma_i}{\left(\gamma_i + \beta^2\right)^3} \left[\mathbf{U}^{\mathbf{T}}\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-\mathbf{1}}\mathbf{d}'\right]_i \left[\mathbf{U}^{-\mathbf{1}}\mathbf{H}^{-\mathbf{1}}\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\mathbf{d}}^{-\mathbf{1}}\mathbf{d}'\right]_i \right] \tag{A6}$$

Thus, $\beta = 0$ is only a probability maximum if

$$\sum_{i=1}^{M} \left[\frac{\omega_i}{\alpha^2} - \frac{1}{\gamma_i} \right] < 0 \,. \tag{A7}$$